TURING 图灵数学·统计学丛书

Probability in Electrical Engineering and Computer Science
An Application-Driven Course

EECS 应用概率论

[美] Jean Walrand 著
黄隆波 译

以PageRank等热门的计算机和通信算法为实例，
带你进入概率论的奇妙世界

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

数字版权声明

Jean Walrand

在美国加州大学伯克利分校取得EECS博士学位，自1982年以来一直在该校任教，研究兴趣包括随机过程、排队论、通信网络、博弈论和互联网的经济性。Walrand教授是比利时－美国教育基金会和IEEE的研究员，曾经荣获兰彻斯特奖、莱斯论文奖、IEEE小林宏治奖和ACM测量与评估专业卓越成就奖。

黄隆波

清华大学交叉信息研究院Tenure-track助理教授，博士生导师，中组部“青年千人计划”入选者。于2011年在美国南加州大学电子工程系获得博士学位，于2011年到2012年在美国加州大学伯克利分校电子工程与计算机科学系担任博士后研究员。在美国麻省理工学院信息与系统决策实验室（LIDS）、法国贝尔实验室与香港中文大学网络编码研究所（INC）等机构担任访问学者与访问教授，共发表IEEE/ACM顶级杂志和会议论文40余篇，曾获邀为多个IEEE/ACM顶级期刊审稿并多次担任IEEE/ACM会议程序委员。

图灵数学·统计学丛书

Probability in Electrical Engineering and Computer Science

An Application-Driven Course

EECS 应用概率论

[美] Jean Walrand 著

黄隆波 译

人 民 邮 电 出 版 社

北 京

图书在版编目（CIP）数据

EECS应用概率论 /（美）瓦尔朗（Walrand,J.）著；
黄隆波译. -- 北京 : 人民邮电出版社, 2015.9
（图灵数学·统计学丛书）
ISBN 978-7-115-39896-3

Ⅰ. ①E… Ⅱ. ①瓦… ②黄… Ⅲ. ①概率论 Ⅳ.
①O211

中国版本图书馆CIP数据核字(2015)第162047号

内 容 提 要

本书精心选取了6个当前热门的科技应用：谷歌 PageRank 算法、链路复用技术、数字链路通信、追踪预测、语音识别和路线规划，并通过讲述概率论在不同应用中的作用来详细介绍基础的概率知识以及概率论中的重要概念，包括马尔可夫链、大数定律、中心极限定理、假设检验、最小方差预测等。

读者对象包括高等院校电子、计算机和统计类专业的高年级本科生及研究生，也包括从事算法、语音识别和机器学习等工作的软件工程师和通信工程师等。

◆ 著　　　[美] Jean Walrand
译　　　黄隆波
责任编辑　朱　巍
执行编辑　杨　琳
责任印制　杨林杰

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路11号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京　　　　印刷

◆ 开本：700×1000　1/16
印张：18　　彩插：4
字数：352千字　　2015年9月第1版
印数：1－4 000册　　2015年9月北京第1次印刷

著作权合同登记号　图字：01-2014-4186号

定价：69.00元

读者服务热线：(010)51095186转600　印装质量热线：(010)81055316

反盗版热线：(010)81055315

广告经营许可证：京崇工商广字第0021号

版 权 声 明

译　者　序

你现在看到的这本概率书，和市面上绝大部分的概率书籍都非常不同.

尽管概率论在工程与自然科学中已经有了广泛的应用，但现有的绝大部分概率书籍往往沿用数学教材的经典方式，即从定义出发、列举重要结果定理证明，并给出相关例子. 虽然这一模式极为有效，但是数学教材中的例子以说明概念为主，鲜少阐述概率论如何在实际的工程问题中发挥作用. 因此，许多国内外 EECS（Electrical Engineering and Computer Science，电子工程与计算机科学系）的同学在学习概率论的时候，除了觉得好玩有趣之外，往往容易感到困惑：不知道自己为什么需要学习概率论，也不知道概率论究竟能够应用在哪些地方. 正因如此，不少同学认为概率论是一门只有做学术研究的人才需要的课程.

本书在内容编撰上另辟蹊径，从应用出发，通过精心选取 EECS 领域核心方向中的几个基本问题，深入浅出却不失系统性地介绍概率论的基础及其在实际问题中的使用. 这其中包括：

(1) PageRank 与马尔可夫链

(2) 网络多路复用与大数定律和中心极限定理

(3) 数字链路与编码和假设检验

(4) 追踪定位与最小方差估计和卡尔曼滤波器

(5) 语音识别、大数据与隐马尔可夫链和期望最大化

(6) 路线规划与马尔可夫决策问题和线性二次型高斯问题

以上列举的应用问题通常很少与书中涵盖的概率论知识一同出现. 本书作者这样安排，正是希望通过一种以应用为主的方式，解答读者在概率论学习过程中对其实际效用的疑惑；也希望通过这种讲授方式，使读者对概率论的重要性有更好的认识；并且希望读者在学习工作中遇到问题时，能从概率论的角度出发进行思考.

本书作者 Jean Walrand 教授在加州大学伯克利分校的电子计算机系从教 30 余年，是计算机网络、随机过程与控制研究的大师. 他不仅在概率论与工程研究上有极高的造诣，而且深谙概率工具在 EECS 领域中应用的关键. 在此书中，Walrand 教授无疑成功地将这两者进行了结合.

在著书之前，Walrand 教授就已在加州大学伯克利分校以基于应用模式讲授概率论的方式积累了多个学期的课堂讲学经验，并根据学生与教师的反馈不断改进此书的内容选取与编排. 因此，这无疑是 Walrand 教授精雕细琢的倾力之作. 译者从事随机建模与分析多年并曾师从 Walrand 教授，但在翻译的过程中，亦不时被作者精彩新颖的观点与讲解打动，受益匪浅. 译者希望，译本也能为读者带来同样的感受.

最后，特别致谢清华大学电子系的连婧与别致两位同学. 她们在本书的翻译、排版与审校过程中给予译者鼎力协助. 同时也特别感谢图灵文化的朱巍女士和杨琳女士在本书的编辑与审校中给予的大力支持. 正因有她们的努力，本书的翻译工作才得以圆满完成.

黄隆波

2015 年 6 月 29 日

前　言

这是一本介绍概率论在电子工程和计算机科学领域应用的书．本书关注一些具有代表性并用到了大量概率论知识的重要实际科技应用．我们无意涵盖所有的应用，一来内容过于宽泛，二来工程量过于浩大．

如同微积分和离散数学一样，掌握概率建模和分析对于计算机科学家和电子工程师来说至关重要．由于这些科学家和工程师应用和设计的复杂系统通常在动态的环境中运作，理解和量化不确定性对系统产生的影响已成为系统设计过程中极为重要的一环．

本书是为加州大学伯克利分校电子工程和计算机科学系的 EECS126 概率论课程而编写的．该课程是为大三、大四学生开设的进阶课程．这门课的学生大多已选修过基础的概率论课程．他们了解事件、概率、条件概率、贝叶斯公式、离散随机变量及期望等概念，并且对矩阵运算有基本的了解（这些概念在附录中都有所回顾）．这些学生都非常聪明、勤奋并对复杂的新知识具有很大兴趣．在这门课程上，学生会学习并了解马尔可夫链、随机动态规划、检验和估计等知识．他们不仅能直观理解这些概念和方法，同时也能熟知如何运用它们．

目前绝大部分概率论的入门书籍在内容介绍上仍然延用了概率空间、随机变量、数学期望、检验、估计和马尔可夫链的顺序．相比之下，本书内容编排的一大特点是以应用作为导引．对于概率论中的每一个知识点，本书均以一个电子及计算机领域相应的重要实际应用作为媒介，详细讲述该知识点背后的理论及其如何在应用中发挥功用．我们认为，阐述理论在实际应用中功用的讲授方式能让人们更容易了解到概率论的巨大价值．不仅如此，本书同时强调让学生独立应用 Matlab 和 Simulink 数学软件完成对课程项目的仿真和计算．这些题目经过精心设计，旨在加强学生对概念的直观理解，并为其日后进行独立探索打好基础．除了最后一章和附录之外，本书中其余的章节均分为 A 和 B 两部分．A 部分介绍主要的概念，而 B 部分则涵盖知识点中更为深奥的部分．因此，本书的一种使用方式是为大三的学生讲授 A 部分和附录中的内容；另一种授课方式是开设一门两学期的课程，在第一阶段课程中讲授 A 部分，将 B 部分纳入第二阶段的学习中．如果想要提升课程难度，也可以在一门课程中先教授 A 部分，再教授 B 部分．按章节顺序授课也不失为一种不错的选择．本书的最后一章收录

了一些前沿课题，读者和教师可以有选择地使用.

附录部分为大多数读者提供了有用的背景知识. 附录 A 回顾了概率论的基本概念. 根据学生背景知识的不同，教师可以选择从附录 A 的综述讲起. 附录 B 复习了一些线性代数的基本知识，而附录 C 提供了一些 Matlab 的实例. 采用 Matlab 是因为现在许多大学都为学生提供了购买 Matlab 的许可. 即使学校没有获得授权，学生也能负担起该软件的学生版. 学生通过在 Matlab 中尝试一些算法并模拟一些系统，能更好地理解相关知识点和认知系统的具体运作方式. 不仅如此，让学生轻松实现一些看起来复杂的算法会让他们在学习中获得更多的成就感和乐趣.

书中的理论从基于未知量的*模型*入手. 我们将这样的未知量记为 X 和 Y，并考虑一个能用于计算关于函数 $h(X)$期望值 $E(h(X))$的模型. 举例来说，X 可表示一块太阳能电池板在一个月中每天产生的能量，$h(X)$为这块电池板一个月产生的总能量. 那么，$E(h(X))$就是电池板每月平均产生的能量. 其他例子还包括通信网络中的平均数据包延时，数据中心完成一件工作的平均耗时，等等.

估计 $E(h(X))$的过程称为*性能评估*. 在许多情况下，在需要处理未知量的系统中都存在一些参数 θ. 人们可以通过调节这些参数的值来调整系统作业. 比如说，太阳能电池板的朝向可以进行调整，数据中心的运作也能调整. 在这种情况下，我们可以通过将系统性能描述成未知量 X 和可变参数 θ 的函数 $h(X,\theta)$来衡量参数取值对系统的影响. 如此一来，优化系统这一问题就等同于寻找能最大化 $E(h(X,\theta))$的参数值 θ. 这个问题通常并不简单，尤其是在 $E(h(X,\theta))$没有解析表达式时. 本书会讲解这样的*优化问题*.

$$X \overset{?}{\rightarrow} E(h(X)) \qquad \max_{\theta} E(h(X,\theta))$$

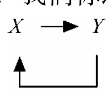
评估　　　　　　优化

在许多问题中，我们能观测到 Y，但是希望估计未被观测的 X 值. 举个例子，X 可以是发射机发出的信号，而 Y 是接收机收到的信号. 根据 Y 估计 X 值的问题是一个*推断问题*. 这方面的例子包括探测问题（附近有没有火源，你有没有感冒）和估计问题（发出 GPS 信号的 iPhone 在哪里）.

$$Y \overset{?}{\rightarrow} X$$

推断

最后，有一类问题关注基于观测的动态系统控制. 比如说，一辆利用激光测距仪、GPS 和相机进行观测的无人驾驶汽车. 我们称之为*控制问题*.



控制

本课程讨论性能评估、优化、推断和控制问题．这其中的一些问题在计算机科学中被称为人工智能，而在电子工程中则被称作统计信号处理．概率学家称之为样本，而数学家将其称为特例．我们通过讲述具体的实际应用，如网络搜索、多路传输、数字通信、语音识别、图像跟踪、路径规划和推荐系统等来介绍这些技术．在学习和了解这些概念的过程中，我们也会为大家介绍这个领域的相关巨擘．

本书的内容十分有趣、能引人思考．希望你们能分享我对这些想法的热情．

十分感谢在本书撰写过程中给予我大力支持的各位同事和学生们．我要特别感谢仔细阅读了手稿的黄隆波博士和 Ramtin Pedarsani，感谢提供了宝贵意见的 Abhay Parekh 博士、David Aldous 教授、Venkat Anantharam 教授、Tom Courtade 教授、Michael Lustig 教授、John Musacchio 教授、Kannan Ramchandran 教授、Anant Sahai 教授、David Tse 教授、Martin Wainwright 教授和 Avideh Zakhor 教授．感谢担任课程助教并设计了课程作业的 Stephan Adams、Vijay Kamble、Shiang Jiang 博士、Sudeep Kamath 博士、Jerome Thai、Baosen Zhang 博士和 Antonis Dimakis 教授．感谢教授我概率论的 Pravin Varaiya 教授和 Eugene Wong 教授，以及支持我的 Tsu-Jae King Liu 教授．最后，感谢提供反馈的 EECS126 课程的学生们．

网站 https://sites.google.com/site/walrandpeecs/home 为本书提供了额外的资源，如勘误、附加题、评论和幻灯片等．教师也可以使用该网站来寻求问题的解决方案．

目 录

第1章

PageRank—A

应用：在网页搜索中，按网页相关度由高到低进行排序
主题：有限离散时间马尔可夫链，强大数定律

背景知识：附录 A.1 ~ A.2

搜索引擎采用不同的算法将网页按给定的关键字以相关度递减的方式排序．其中一种算法的思想是计算马尔可夫链的稳态分布．本章讨论这种分布的存在性与唯一性，以及在随机浏览时找到一个特定网页所需的平均时间．我们将采用强大数定律证明马尔可夫链处于每个特定状态的时间比例是收敛的．

1.1 模型

互联网由一系列相互链接的网页组成．这些网页和它们之间的链接关系构成了一张图．如图 1-1 所示，图中的节点是所有的网页 $\mathcal{X}$．若网页 i 有一个到网页 j 的链接，则图中有一条由 i 到 j 的弧（有向边）．

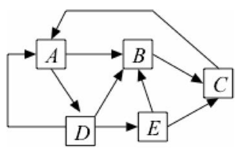


图 1-1　在网络中，网页总是指向其他的网页．在本图中 $P(A, B) = 1/2$，$P(D, E) = 1/3$

直观上看，一个高级别网页所指向的网页也应具有较高的级别（在实际中，除了我们在此讨论的网页级别度量方式，搜索引擎的排序结果还取决于网页中关键字是否出现以及其他许多因素）．因此，我们定义网页 i 的级别 $\pi(i)$为一个正数，并且满足

$$\pi(i) = \sum_{j \in \mathcal{X}} \pi(j) P(j,i), i \in \mathcal{X}.$$

$P(j, i)$表示所有网页 j 的外向链接中指向 i 的链接所占的比例．如果 j 没有指向 i 的链接，则 $P(j, i) = 0$．在上述例子中，我们有 $P(A, B) = 1/2$，$P(D, E) = 1/3$，$P(B, A) = 0$ 等．这一算法的思想源自拉里・佩奇（Larry Page，见图 1-2），这也是 PageRank（佩奇排序）这一名字的由来．



图 1-2　谷歌公司创始人之一拉里・佩奇

可以将这些等式以矩阵的形式记作

$$\pi = \pi P . \tag{1.1}$$

这里，π 是一个以 $\pi(i)$为分量的行向量，而 P 是一个以 $P(i, j)$为元素的方阵．

式（1.1）称为*平衡方程*．这里我们注意到，如果一个向量 π 是这个方程的解，那么 π 的任意倍数也是这个方程的解．为方便起见，我们将解归一化，使得所有网页的级别之和为 1：

$$\sum_{i \in \mathcal{X}} \pi(i) = 1 . \tag{1.2}$$

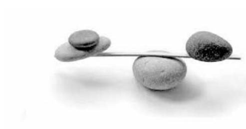


图 1-3　这是平衡方程吗

对于图 1-1 中的例子，平衡方程为

$$\begin{aligned}
\pi(A) &= \pi(C) + \pi(D)(1/3) \\
\pi(B) &= \pi(A)(1/2) + \pi(D)(1/3) + \pi(E)(1/2) \\
\pi(C) &= \pi(B) + \pi(E)(1/2) \\
\pi(D) &= \pi(A)(1/2) \\
\pi(E) &= \pi(D)(1/3).
\end{aligned}$$

再加上 $\pi(i)$的和为 1 这一条件，可以得到

$$\pi = [\pi(A), \pi(B), \pi(C), \pi(D), \pi(E)] = \frac{1}{39}[12, 9, 10, 6, 2].$$

由此可以看到，网页 A 具有最高的级别，网页 E 具有最低的级别. 运用这一方法的搜索引擎会将这些网页级别与其他因素相结合来进行网页排序. 一些搜索引擎也会采用这一度量的变种来对网页进行排序.

1.2　马尔可夫链

想象一下你正在浏览网页：假设你在网页 i 上浏览了一个单位时间，然后随机点击进入了网页 i 指向的一个网页. 在这个过程中，从网页 i 到网页 j 的概率正好为 $P(i, j)$，与前面的例子相同.

1.2.1　定义

考虑一个包含节点 $\mathcal{X} = \{1, 2, \cdots, N\}$ 和有向边的有限图. 假设其中有些节点具有指向自己的边. 图中每条边(i, j)，都有一个权重 $P(i, j) > 0$. 这些权值使得每个点外向边的权和为 1. 根据习惯，如果图中没有从 i 到 j 的边，则 $P(i, j)$为 0.

以上述方式得到的矩阵 $P = P[(i, j)]$叫作随机矩阵. 这种矩阵的每个元素均为非负，并且每行的和为 1. 现在，我们定义以下过程$\{X(n), n \geqslant 0\}$：在时刻 0 的时候用 $X(0)$表示系统所处的状态；随后的每一个时刻 n，系统由状态 $X(n-1) = i$ 跳到状态 $X(n) = j$ 的概率为 $P(i, j)$，即系统所处在的状态仅由 $X(n-1)$与 P 决定. 这样定义的过程$\{X(n), n \geqslant 0\}$被称为马尔可夫链. 该有限图则被称为马尔可夫链的状态转移图.



图 1-4　安德雷 · 马尔可夫（1856—1922）

图 1-5 展示了 3 个马尔可夫链的状态转移图.

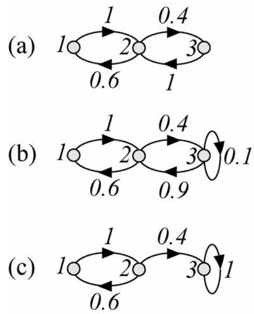


图 1-5　3 个具有不同转移概率的 3 状态马尔可夫链（状态为 {1, 2, 3}）

上面的描述可以用以下公式来表述：

$$P[X(n+1)=j \mid X(n)=i, X(m), m<n]=P(i,j), \forall i,j \in \mathcal{X}, n \geqslant 0. \qquad (1.3)$$

从状态 i 移动到状态 j 的概率与先前状态无关．这种“健忘”的性质叫作马尔可夫性质．这也正是 $X(n)$被称为“状态”的原因：它包含了所有预测该过程未来状态所需要的相关信息．

1.2.2　n步后的分布和稳态分布

C.3

假如马尔可夫链在第 n 步（$n\geqslant 0$）处于状态 j 的概率为 $\pi_n(j)$，则它在第 n+1 步处于状态 i 的概率 $\pi_{n+1}(i)$可由下面的公式得到

$$\pi_{n+1}(i)=\sum_{j\in\mathcal{X}}\pi_n(j)P(j,i), i\in\mathcal{X}. \qquad (1.4)$$

事实上，马尔可夫链在第 n+1 步处于状态 i 的事件可以表示为马尔可夫链在第 n 步处于状态 j，但在第 n+1 步时处于 i 事件的集合．由于这些事件为互斥事件（即每次只能处在一个状态 j），联合事件发生的概率正好为各个时间的概率之和，且马尔可夫链在第 n 步处于状态 j 而第 n+1 步处于状态 i 的概率是 $\pi_n(j)P(j,i)$.

在矩阵记法下，

$$\pi_{n+1}=\pi_n P,$$

所以，

$$\pi_n=\pi_0 P^n, n\geqslant 0. \qquad (1.5)$$

这里我们注意到，对于 $n\geqslant 0$ 和 $i\in\mathcal{X}$的所有情况，当且仅当 π_0 满足式（1.1）的平衡方程时有 $\pi_n(i)=\pi_0(i)$．在这种情况下，称 π_0 为稳态分布．因此，稳态分布就是式（1.1）的非负解 π，并满足分量之和为 1.

1.3 分析

经过上述分析与推导之后，我们会很自然地提出以下的问题.

Q1：是否每个马尔可夫链都具有一个稳态分布？

Q2：该稳态分布是否唯一？

Q3：π_n 是否总是趋向于稳态分布？

1.3.1 不可约性和非周期性

为了回答上面的三个问题，首先定义马尔可夫链的一些性质.

定义 1.1 不可约的，非周期性的，周期性的

(a) 如果一个马尔可夫链可以从一个状态转移到任意其他状态（也许经过许多步的跳转），那么该马尔可夫链是**不可约的**.

(b) 假设一个马尔可夫链是不可约的，并且定义[①]

$$d(i) := g.c.d\{n \geqslant 1 \mid P^n(i,i) > 0\}. \tag{1.6}$$

所有 i 都有相同的值 $d(i) = d$（如引理 2.8 所示）. 如果 $d = 1$，那么该马尔可夫链是**非周期性的**；否则它是**周期性的**，且周期为 d. □

在图 1-5 中，马尔可夫链(a)和(b)是不可约的，而(c)是可约的. 同时，(a)是周期性的，而(b)是非周期性的.

1.3.2 大数定律

我们可以通过一些简单的例子说明 Q2、Q3 的答案并非总是肯定的. 比如，对于一个没有状态转移的马尔可夫链来说，每一个分布都是它的稳态分布. 再比如，考虑一个只在状态 0、1 之间来回转变的马尔可夫链. 如果从 $\pi_0(0) = 1$ 开始，则它在偶数时刻 $\pi_n(0) = 1$，在奇数时刻 $\pi_n(0) = 0$. 因此 π_n 并不收敛.

尽管有这些反例，我们仍然能得到下面的重要结论.

定理 1.2 有限状态马尔可夫链大定律

(a) 每一个不可约的有限马尔可夫链都拥有唯一的稳态分布 π，并且 $\pi(i)$ 表示马尔可夫链 $X(n) = i$ 的长期时间比例.

(b) 如果这个马尔可夫链是非周期性的，其 n 步分布 π_n 随着 n 收敛到 π. □

---

① 如果 S 是正整数集，$g.c.d.(S)$表示这些正整数的最大公约数.

在以上的定理中，$X(n) = i$ 的长期时间比例定义如下[①]：

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}1\{X(n)=i\}.$$

在这个公式里，当 $X(n) = i$ 时 $1\{X(n)\}$取 1，否则取 0. 一般来说，如果 A 是一个事件，那么当该事件发生时 $1\{A\}$取 1，否则取 0. 我们称 $1\{A\}$是事件 A 的指示函数.

这个定理说明，如果马尔可夫链是不可约的，上述极限存在并且等于 $\pi(i)$，而且该极限不依赖于随机变量的具体取值. 这意味着每次模拟该马尔可夫链都会得到相同的极限（本章练习 8 将验证这一点）.

1.3.3 长期时间比例

为什么马尔可夫链在一个状态的时间比例收敛？在浏览网页的例子中，如果统计花费在网页 A 上的时间，将其除以花费在所有网页上的时间 n，会发现这个值随着 n 收敛到 $\pi(A)$.

这一结果和我们熟知的抛硬币试验类似：如果重复抛掷一枚均匀的硬币，正面朝上的频率会收敛于 50%. 因此，尽管硬币没有任何记忆，但是它却保证了正面朝上的概率为 50%！这是为什么呢？

事实上，这些收敛结果都可以被视为大数定理的特例. 大数定理对于直观地理解概率和统计规律具有极核心的作用. 正是因为它，我们能对不确定的结果进行预测. 以下是大数定律的具体表述. 我们将在第 2 章进行更具体的讨论.

定理 1.3 强大数定律

如果$\{X(n)\geqslant 1\}$是一系列均值为 μ 的独立同分布的随机变量. 那么当 $n\to\infty$ 时，$\dfrac{X(1)+\cdots+X(n)}{n}\to\mu$ 的概率为 1.

因此，样本均值 $Y(n) := (X(1)+\cdots+X(n))/n$ 以概率 1 收敛到期望值（见图 1-6）. 这里的样本均值 $Y(n)$是随机变量：对于每个 n，$Y(n)$的值取决于随机变量 $X(m)$的具体取值. 因此，重复这一试验可能会得到不同的值. 然而，它的极限值以概率 1 收敛到 μ. 这种收敛模式名为几乎处处收敛.

① 求和表示起初 N 步内马尔可夫链在状态 i 的总时间. 除以 N 表示时间比例，取极限得出长期时间比例.

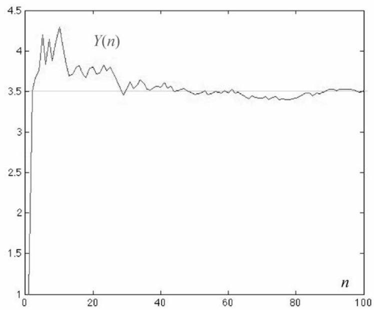


图 1-6　当抛掷一枚均匀的骰子时，样本均值收敛于 3.5

1.4　击中时间

在图 1-1 中，假设从网页 A 开始浏览．在每一步以相同的概率点击进入当前页的一个外部链接网页，那么需要多少步才能到达网页 E？我们把这个时间叫作网页 E 的击中时间，也叫首通时间，记作 T_E．从图中可以看到，T_E 的最小值为 2．当然，T_E 也有可能比 2 大得多.



图 1-7　这可不是我们说的击中时间

1.4.1　平均击中时间

我们的目标是计算从 $X_0 = A$ 开始的 T_E 的均值：

$$\beta(A) := E[T_E \mid X_0 = A].$$

完成这一计算的关键在于计算从所有可能的初始页面到 E 的平均击中时间．也就是说，我们要计算当 $i = A, B, C, D, E$ 时的 $\beta(i)$，

$$\beta(i) := E[T_E \mid X_0 = i].$$

这么做的原因在于，从 A 开始命中 E 的平均时间与从 B 和从 D 开始的平均击中时间相关联，而它们又和从 C 开始的平均击中时间有关．首先得到

$$\beta(A)=1+\frac{1}{2}\beta(B)+\frac{1}{2}\beta(D). \qquad (1.7)$$

可以这样理解这个式子：从 A 开始经过 1 步后，马尔可夫链以 1/2 的概率转到状态 B，而以 1/2 的概率处于状态 D．因此，在 1 步后，马尔可夫链命中 E 的平均时间以 1/2 的概率等于从 B 开始到命中 E 的平均时间，以 1/2 的概率等于从 D 开始到命中 E 的平均时间．

这个情况和下面这个例子类似：抛掷一枚均匀硬币，如果正面朝上可以得到随机的 X 元，如果反面朝上则可以得到随机的 Y 元．平均下来，可以得到

$$\frac{1}{2}E(X)+\frac{1}{2}E(Y).$$

通过类似的论证，可以得到以下和式（1.7）并列的等式：

$$\begin{aligned}
&\beta(B)=1+\beta(C)\\
&\beta(C)=1+\beta(A)\\
&\beta(D)=1+\frac{1}{3}\beta(A)+\frac{1}{3}\beta(B)+\frac{1}{3}\beta(E)\\
&\beta(E)=0.
\end{aligned}$$

这些方程和式（1.7）一起构成了*首步方程*（First Step Equation，FSE）．通过求解方程得到

$$\beta(A)=17, \beta(B)=19, \beta(C)=18, \beta(D)=13, \beta(E)=0.$$

1.4.2 击中另一状态之前命中某一状态的概率

仍然考虑前面的例子，但是这次考虑从网页 A 开始先访问 C 再访问 E 的概率．我们将这一概率记作

$$\alpha(A)=P[T_C<T_E \mid X_0=A].$$

和前面的情况类似，需要计算所有 $i=A, B, C, D, E$ 时的 $\alpha(i)$．首先

$$\alpha(A)=\frac{1}{2}\alpha(B)+\frac{1}{2}\alpha(D). \qquad (1.8)$$

式（1.8）成立是因为，从 A 开始经过 1 步之后会有两种情况：第一，以 1/2 的概率处于状态 B，然后会以 $\alpha(B)$的概率在 E 之前访问 C；第二，以 1/2 的概率处于状态 D，然后会以 $\alpha(D)$的概率在 E 之前访问 C．因此，从 A 开始在 E 之前访问 C 的事件是这两个互斥事件的集合，因为要么先经过 B，要么先经过 D，然后再在 E 之前访问 C．将这两个事件的概率相加，得到式（1.8）．

如同计算平均击中时间一样，我们也得到以下的等式组：

$$
\begin{aligned}
&\alpha(B)=\alpha(C)\\
&\alpha(C)=1\\
&\alpha(D)=\frac{1}{3}\alpha(A)+\frac{1}{3}\alpha(B)+\frac{1}{3}\alpha(E)\\
&\alpha(E)=0.
\end{aligned}
$$

这些方程和式（1.8）一起叫作首步方程．解之可得

$$\alpha(A)=\frac{4}{5}, \alpha(B)=1, \alpha(C)=1, \alpha(D)=\frac{3}{5}, \alpha(E)=0 .$$

1.4.3 马尔可夫链的首步方程

现在把前面例子中得到的结果推广到任意的有限马尔可夫链．记 $\mathcal{X}=\{1,2,\cdots,N\}$ 为状态集合，并记 P 为马尔可夫链的转移概率矩阵．此时，定义 T_i 为状态 i 的击中时间．对于状态集合 $A\subset\mathcal{X}$，定义 $T_A=\min\{n\geqslant 0|X(n)\in A\}$ 为集合 A 的击中时间．

第一个扩展考虑 T_A 的平均值．定义

$$\beta(i)=E[T_A \mid X_0=i], i\in\mathcal{X},$$

则首步方程是

$$\beta(i)=\begin{cases}1+\Sigma_j P(i,j)\beta(j), & i\notin A\\ 0, & i\in A.\end{cases}$$

接下来，对两个不相交的集合 A 和 B，即 $A,B\subset\mathcal{X}$ 且 $A\cap B=\varnothing$，考虑在击中集合 B 之前击中集合 A 的概率．定义

$$\alpha(i)=P[T_A<T_B \mid X_0=i], i\in\mathcal{X},$$

然后可以得到下述的首步方程组

$$\alpha(i)=\begin{cases}\Sigma_j P(i,j)\alpha(j), & i\notin A\cup B\\ 1, & i\in A\\ 0, & i\in B.\end{cases}$$

第三个扩展是

$$Y=\sum_{n=0}^{T_A} h(X(n)) .$$

这可以理解为每次访问状态 i 时我们得到 $h(i)$的奖励，直到马尔可夫链进入状态集合 A 为止．令

$$\gamma(i)=E[Y \mid X_0=i], i\in\mathcal{X},$$

可以得到以下的首步方程

$$\gamma(i)=\begin{cases}h(i)+\Sigma_j P(i,j)\gamma(j), & i\notin A\\ h(i), & i\in A.\end{cases} \tag{1.9}$$

第四个扩展是研究

$$Z=\sum_{n=0}^{T_A}\beta^n h(X(n))\text{ 的值.}$$

这里 β 为折扣因子．令

$$\delta(i)=E[Z\mid X_0=i],$$

则首步方程是

$$\delta(i)=\begin{cases}h(i)+\beta\Sigma_j P(i,j)\delta(j), & i\notin A\\ h(i), & i\in A.\end{cases}$$

希望以上这些例子能让读者大致认识到，许多关于有限状态马尔可夫链的问题都能找到解答的方式．这是一件十分幸运的事情，因为马尔可夫链在对工程和自然系统的建模与模拟中都有着广泛的应用．

1.5 小结

- 马尔可夫链：状态，转移概率，不可约的，非周期性的，稳态分布，击中时间
- 强大数定律
- 大数定律：不可约意味着有唯一的稳态分布，这个分布等于长期时间比例；如果既不可约又是非周期性的，则收敛于稳态分布
- 击中时间：首步方程

重要方程与公式

马尔可夫链的定义	$P[X(n+1)=j\mid X(n)=i,X(m),m<n]=P(i,j)$	式(1.3)
马尔可夫链$X(n)$的概率方程	$\pi_n=\pi_0P^n$	式(1.5)
平衡方程	$\pi P=\pi$	式(1.1)
首步方程	$\gamma(i)=h(i)+\Sigma_j P(i,j)\gamma(j)$	式(1.9)

1.6 参考资料

关于马尔可夫链有非常多优秀的书籍，我最喜欢的是 Grimmett 和 Stirzaker 的 *Probability and Random Processes*，以及 Bertsekas 和 Tsitsiklis 的 *Introduction to Probability*．最初介绍 PageRank 的专利文献是“Method for node ranking in a linked

database”. Easley 和 Kleinberg 的电子书 *Networks，Crowds，and Markets: Reasoning About a Highly Connected World* 对社交网络的探讨很有启发性. 该书的第 14 章讨论了 PageRank.

1.7 练习

1. 构造一个不可约的马尔可夫链，要求其分布收敛于唯一的稳态分布.

2. 给出一个马尔可夫链，要求其分布的极限收敛且依赖于初始分布.

3. 你能找到一个分布不收敛、不可约、非周期性的有限马尔可夫链吗?

4. 给出一个有限、不可约、非周期性的马尔可夫链，要求它以很慢的速度收敛于稳态分布.

5. 证明：如果 $X(n)$是一个马尔可夫链，则函数 $Y(n) = g(X(n))$ 有可能不是一个马尔可夫链.

6. 如果一个马尔可夫链是独立同分布的随机变量序列，它是否是不可约的和非周期性的?

7. 考虑马尔可夫链 $X(n)$，其状态如图 1-8 所示，其中 $a, b\in(0,1)$.

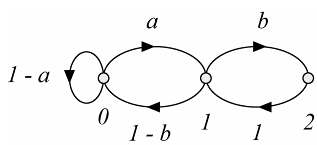


图 1-8　第 7 题的马尔可夫链

(a) 证明这个马尔可夫链是非周期性的.

(b) 计算 $P[X(1) = 1, X(2) = 0, X(3) = 0, X(4) = 1|X(0) = 0]$.

(c) 计算其稳态分布.

(d) 令 $T_i = \min\{n\geqslant 0|X(n) = i\}$，计算 $E[T_2|X(0) = 1]$.

8. 用 Matlab 写一个 K 状态马尔可夫链$\{X(n), n\geqslant 1\}$的模拟程序. 假设其初始分布为 π，转移概率矩阵为 P. 这个程序需要完成以下任务：

C.3

(a) 绘制$\{X(n), n = 1, \cdots, N\}$;

(b) 绘制 $X(n)$在某些指定状态上所花的时间比例，并将其表示为 m 的函数（$m = 1, \cdots, N$）;

(c) 绘制 $X(n)$等于某些确定状态的概率，其中 $n = 1, \cdots, N$;

(d) 用这个程序模拟一个 5 状态的周期性马尔可夫链;

(e) 用这个程序模拟一个 5 状态的非周期性马尔可夫链.

9. 用你在问题 8 中写的程序模拟图 1-1 和图 1-5 的马尔可夫链.

10. 计算图 1-5 所示的马尔可夫链的稳态分布.

11. 计算图 1-5 所示的马尔可夫链的 $d(1)$、$d(2)$和 $d(3)$（定义在式（1.6）中）.

12. 计算图 1-1 所示的马尔可夫链的 $d(A)$（定义在式（1.6）中）.

13. 设$\{X_n, n\geqslant 0\}$为一个有限状态马尔可夫链. 假设它有唯一的稳态分布 π 且 π_n 对于任意初始分布 π_0 都收敛于 π. 请问以下哪些选项是正确的?

❑ X_n 是不可约的.

❑ X_n 是周期性的.

❑ X_n 是非周期性的.

❑ X_n 可能不是不可约的.

14. 考虑在$\{0, 1\}$上的马尔可夫链$\{X_n, n\geqslant 0\}$. 设 $P(0, 1) = 0.1$，$P(1, 0) = 0.3$，以下哪些选项是正确的?

❑ 该马尔可夫链的样本空间是$\{0,1\}$.

❑ 该马尔可夫链的稳态分布为[0.75, 0.25].

❑ 令 $T_1 = \min\{n\geqslant 0|X_n = 1\}$，那么 $E[T_1|X_0 = 0] = 1.2$.

❑ $E[X_1+X_2|X_0 = 0] = 0.8$.

15. 考虑一个马尔可夫链，其状态转移如图 1-9 所示.

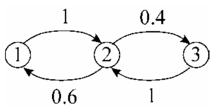


图 1-9　第 15 题的马尔可夫链

(a) 该马尔可夫链的周期是多少? 解释原因.

(b) 找出该马尔可夫链的所有稳态分布.

(c) X_n 的分布 π_n 在 $n\to\infty$时是否收敛? 解释原因.

(d) 马尔可夫链花在不同状态中的时间比例是否收敛? 如果收敛，其极限是多少?

16. 考虑图 1-10 所示的马尔可夫链.

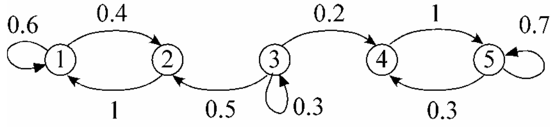


图 1-10　第 16 题的马尔可夫链

(a) 找出该马尔可夫链的所有稳态分布.

(b) 假设 $\pi_0(3) = 1$，求 $\lim_{n\to\infty}\pi_n$.

17. 考虑图 1-11 所示的马尔可夫链.

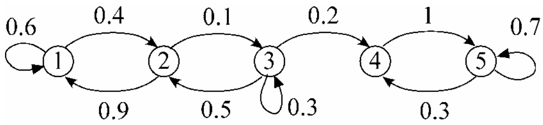


图 1-11 第 17 题的马尔可夫链

(a) 找出该马尔可夫链的所有稳态分布.

(b) 当 $n\to\infty$时，π_n是否收敛？如果收敛，加以证明.

(c) 马尔可夫链花在所有状态中的时间比例是否收敛？并证明之.

18. 考虑一个马尔可夫链，其状态转移如图 1-12 所示.

C.3

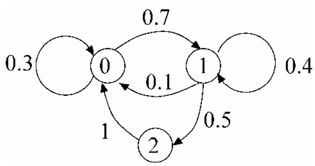


图 1-12 第 18 题的马尔可夫链

(a) 找出该马尔可夫链的稳态分布 π.

(b) 计算从状态 0 到 2 的所需时间期望值.

(c) 用 Matlab 绘制从 0 开始经过 n 步后，马尔可夫链还未到达 2 的概率.

(d) 用 Matlab 对该马尔可夫链进行模拟并绘制在 n 步后花费在不同状态上的时间比例.

(e) 用 Matlab 绘制 π_n.

19. 考虑图 1-13 所示的马尔可夫链$\{X_n, n\geqslant 0\}$，假设 $X_0=0$. 求 X_n在到达 2 之前两次到达 1 的概率.

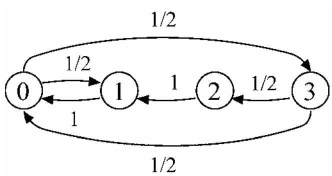


图 1-13 第 19 题的马尔可夫链

20. 画一个 6 状态、不可约、非周期性的马尔可夫链，并且设置转移概率. 在 Matlab 中模拟该马尔可夫链. 画出该链在所有 6 个状态的时间比例. 假设从状态 1 开始，画出它在每个状态的概率.

C.3

21. 重复第 20 题，将条件改为模拟一个周期性的马尔可夫链.

22. 如何使网页排序算法给你的主页一个较高等级？

提示：尝试添加另一个网页并加入一些链接.

23. 证明状态的保持时间是几何分布的.

24. 抛掷一枚骰子，平均来说，要抛掷多少次才能得到最后两次点数和为 10?

25. 抛掷一枚骰子,平均来说,要抛掷多少次才能得到最后三次的点数和至少为 15?

26. 一个二重随机矩阵是非负矩阵，行、列和均为 1. 证明这样一个转移矩阵的稳态分布是均匀的.

27. 假设图 1-5 中的马尔可夫链(c)从状态 1 开始. 计算它在被状态 3 吸收前，访问状态 1 的平均次数.

28. 一个人尝试爬上有 N 级的梯子. 他在爬每一步的时候，都有概率 p 掉回地面，否则爬上一级. 用首步方程的方法从理论上分析他到达顶部所需的平均时间，计算 $N = 1, \cdots, 20$ 和 $p = 0.05, 0.1, 0.2$ 的情况. 利用 Matlab 绘制相应的图像.

C.3

29. 设$\{X_n, n\geqslant 0\}$为一个有限状态、不可约的马尔可夫链,其概率转移矩阵为 P. 证明当 $N\to\infty$ 时，$\frac{1}{N}\sum_{n=0}^{N-1}1\{X_n = i, X_{n+1} = j\}\to\pi(i)P(i,j)$ 的概率为 1.

30. 证明马尔可夫链$\{X_n, n\geqslant 0\}$可以写作 $X_{n+1} = f(X_n, V_n)$，$n\geqslant 0$. 这里的 V_n 是独立于 X_0 的独立同分布的随机变量.

31.设 P 和 $\tilde{P}$ 是随机矩阵，π 是有限集 $\mathcal{X}$ 上的概率分布. 假设

$$\pi(i)P(i,j) = \pi(j)\tilde{P}(j,i), \forall i, j\in\mathcal{X},$$

证明 π 是 P 的稳态分布.

32. 设 X_n 是有限集 $\mathcal{X}$ 上的马尔可夫链. 假设该马尔可夫链的转移图为一棵树，如图 1-14 所示. 证明如果 π 是稳态的，P 是转移矩阵，则它满足如下*细致平衡方程*：

$$\pi(i)P(i,j) = \pi(j)P(j,i), \forall i, j.$$

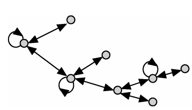


图 1-14　树状的转移图

33. 设$\{X_n, n\geqslant 0\}$是一个在$\{-1,1\}$上的马尔可夫链. 假设 $P(-1,1) = P(1,-1)=a$，其中 a 是给定的且 $a\in(0,1)$. 定义 $Y_n = X_0+\cdots+X_n$，$n\geqslant 0$.

(a) $\{Y_n, n\geqslant 0\}$是马尔可夫链吗? 证明之.

(b) 如何计算

$$E[\tau\mid Y_0 = 1]，其中 \tau = \min\{n>0\mid Y_n = -50 或 Y_n = 30\}?$$

34. 假设无限次地抛掷一枚均匀的硬币. 证明正面出现次数总是比背面出现次数大的概率为 0.

第2章

PageRank—B

主题：样本空间，轨迹；大数定律（弱大数定律、强大数定律）；大定理的证明

背景知识：附录 A.1～A.3，第 1 章

在本章中，我们进一步讨论第 1 章中的一些概念．具体地说，我们会详细解释马尔可夫链的样本空间、大数定律及其证明．

如序言中所述，介绍性的课程可以跳过每章的 B 部分．

2.1 样本空间

首先将马尔可夫链①$X = \{X_n,\ n \geqslant 0\}$与附录 A.1 中的总体框架联系起来．附录 A.1 提到：随机试验通过样本空间来描述，样本空间包含了随机试验的所有可能结果，概率分布建立在样本空间子集合上，以及随机变量是随机试验结果的实值函数．

为了阐明这些概念，首先假设 X_n 为独立同分布的伯努利随机变量，并有 $P(X_n = 1) = P(X_n = 0) = 0.5$．现在考虑一个重复抛掷一枚均匀硬币的实验．这时，这些随机变量可用来描述实验的结果．这个实验的一个可能结果是产生一个全 0 或者全 1 的无穷序列．注意，这里的实验结果并不是指单个的 0 或者 1，而是一个无穷序列，因为结果描述的是一直不停抛掷硬币时发生的情况．因此，样本空间 Ω 是无穷 0、1 序列的集合$\{0, 1\}^\infty$．若 ω 是一个这样的序列，有 $\omega = (\omega_0, \omega_1, \cdots)$，$\omega_n \in \{0, 1\}$．现在，我们可以很自然地定义 $X_n(\omega) = \omega_n$，并将其看作是第 n 次($n \geqslant 0$)抛掷的结果．因此，对所有 $\omega \in \Omega$，都有 $X_n(\omega) \in \Re$，而且每个 X_n 都是一个定义在 Ω 上的实值函数．比如，$X_0(1101001\ldots) = 1$，因为当 $\omega = 1101001\ldots$时，$\omega_0 = 1$．同理，$X_1(1101001\ldots) = 1$，$X_2(1101001\ldots) = 0$．

① 写作 X_n 或 $X(n)$．

为了具体说明随机实验，还需要定义在 Ω 上的概率分布. 最简单的描述方法是对所有 $n \geqslant 0$ 及 $a, b, \cdots, z \in \{0, 1\}$，设

$$P(\{\omega|\omega_0 = a, \omega_1 = b, \cdots, \omega_n = z\}) = P(X_0 = a, \cdots, X_n = z) = 1/2^{n+1}.$$

例如，

$$P(\{\omega|\omega_0 = 1\}) = P(X_0 = 1) = 1/2.$$

同样，

$$P(\{\omega|\omega_0 = 1, \omega_1 = 0\}) = P(X_0 = 1, X_1 = 0) = 1/4.$$

注意我们定义了结果集合或者事件的概率$\{\omega|\omega_0 = a, \omega_1 = b, \cdots, \omega_n = z\}$，而不是每一个特定结果 ω 的概率. 这是因为对于任一特定的 0、1 无限序列，我们观察到其概率为 0. 也就是说，若 $\omega \in \Omega$，则 $P(\{\omega\}) = 0$. 这样的描述并不能告诉我们任何有关抛硬币的信息，比如硬币的偏移以及每一次的抛掷均为独立的. 所以，正确的定义方法是指定事件（结果集合）的概率，而不是独立结果的概率.

每个马尔可夫链都有一个特定的样本空间 Ω，并且每一个 X_n 都是某个结果 ω 的函数 $X_n(\omega)$，其中 ω 取值的集合记为 $\mathcal{X}$. 马尔可夫链的概率分布定义在 Ω 的子集上.

在这个例子中，Ω 可以定义为所有在 $\mathcal{X}$ 中符号的无限序列的集合. 也就是说，$\Omega = \mathcal{X}^{\infty}$，且每一个元素 $\omega \in \Omega$ 均为 $\omega = (\omega_0, \omega_1, \cdots)$，其中 $\omega_n \in \mathcal{X}$，$n \geqslant 0$，如图 2-1 所示. Ω 的选择和抛硬币的样本空间相类似，叫作规范样本空间. 马尔可夫链的一个实际取值序列叫作轨迹，或者马尔可夫链的实现.

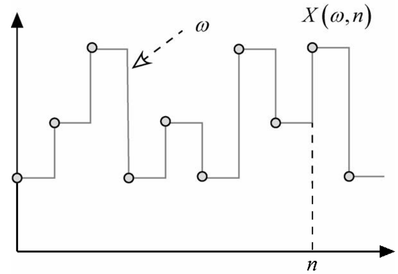


图 2-1　在规范样本空间中，结果 ω 是马尔可夫链的轨迹

接下来需要解决的是如何为 Ω 中的事件分配概率. 采用的方法和抛硬币例子中的相类似，注意到马尔可夫链为某一特定序列的概率为 0（比如全部正面朝上）. 因此，我们为 Ω 的子集而非对单个结果定义概率. 具体地说，对于在 $\mathcal{X}$ 中的 $i_0, i_1, \cdots, i_n$，设对所有 $n \geqslant 0$ 都有

$$P(X_0 = i_0, X_1 = i_1, \cdots, X_n = i_n) = \pi_0(i_0)P(i_0, i_1) \times \cdots \times P(i_{n-1}, i_n). \tag{2.1}$$

这里，$\pi_0(i_0)$是马尔可夫链从状态 i_0 开始的概率，即初始分布.

注意，这一性质与式（1.3）等价．实际上，如果设

$$A_n=\{X_0=i_0, X_1=i_1, \cdots, X_n=i_n\}，以及 A_{n-1}=\{X_0=i_0, X_1=i_1, \cdots, X_{n-1}=i_{n-1}\},$$

那么，由式（1.3）可以得到

$$P(A_n)=P[A_n|A_{n-1}]P(A_{n-1})=P(A_{n-1})P(i_{n-1}, i_n).$$

由数学归纳法进一步得到式（2.1）.

至此，我们已经定义了由马尔可夫链前 n+1 个值所定义的事件的概率．事实上，在 Ω 上存在一个概率分布与所有这些定义值相一致.

2.2 投掷硬币的大数定律

在讨论马尔可夫链之前，先考虑一个简单的例子——抛掷硬币．设$\{X_n, n\geqslant 0\}$是独立同分布的伯努利随机变量，服从 $P(X_n=1)=P(X_n=0)=0.5$．可以认为 $X_n=1$ 表示第 n 次抛掷时正面朝上，$X_n=0$ 代表反面朝上．我们想要证明，如果一直不断地抛掷硬币，那正面朝上的百分数会收敛到 50%．以下是这一结论的两种严格数学表达方式.

2.2.1 依概率收敛

第一种严格的数学表达如下：当 n 很大时，例如 $n=10^5$，正面朝上的比例几乎都是 50%，甚至不会偏离 1%．我们要证明，在 10^5 次抛掷中，正面朝上的百分数大于 51%或小于 49%的可能性很小．并且，只要多抛掷硬币，这一可能性就会尽可能地减小.

下面来证明这一点．定义 $Y_n=\dfrac{X_0+\cdots+X_{n-1}}{n}$，为前 n 次抛掷正面朝上的比例．根据切比雪夫不等式，有：

$$P(|Y_n-E(Y_n)|\geqslant\varepsilon)\leqslant\frac{\operatorname{var}(Y_n)}{\varepsilon^2}. \qquad (2.2)$$



图 2-2 帕夫努季·切比雪夫（1821—1884）

为了理解式（2.2），要注意到[①]

① $1\{C\}$为指示函数．其当 C 为真时取值 1，反之取 0.

$$1\{|Y_n - E(Y_n)| \geqslant \varepsilon\} \leqslant \frac{(Y_n - E(Y_n))^2}{\varepsilon^2}. \quad (2.3)$$

若 $Y_n - E(Y_n) \geqslant \varepsilon$，则 $(Y_n - E(Y_n))^2 \geqslant \varepsilon^2$．因此，如果不等式（2.3）左边为 1，则右边大于等于 1．如果左边为 0，那么左边小于等于右边．所以，式（2.3）总是成立．因为 $E(1_A) = P(A)$并且 $E((X-E(X))^2) = \text{var}(X)$，所以对式（2.3）两边取期望值就能得到式（2.2）.

现在计算 Y_n 的期望与方差．我们可以先得到 $E(Y_n) = 0.5$．注意，如果用 a 乘以一个随机变量，其方差会乘以 a^2．并且，独立随机变量和的方差是它们方差的和．于是得到

$$\text{var}(Y_n) = \frac{\text{var}(X_0 + \cdots + X_{n-1})}{n^2} = \frac{n\,\text{var}(X_0)}{n^2}.$$

所以，

$$P(|Y_n - 0.5| \geqslant \varepsilon) \leqslant \frac{\text{var}(X_0)}{n\varepsilon^2}.$$

又因为 $X_0 = B(0.5)$，所以有

$$\text{var}(X_0) = E(X_0^2) - (E(X_0))^2 = E(X_0) - (E(X_0))^2 = 0.5 - 0.25 = 0.25.$$

由此得到

$$P(|Y_n - 0.5| \geqslant \varepsilon) \leqslant \frac{1}{4n\varepsilon^2}.$$

这里如果选择 $\varepsilon = 1\% = 0.01$，由上式可得，当 $n = 10^5$ 时，

$$P(|Y_n - 0.5| \geqslant 1\%) \leqslant \frac{2500}{n} = 0.025.$$

事实上，从上面的推导可以看出，当 n 趋向无穷的时候，$P(|Y_n - 0.5| \geqslant \varepsilon) \to 0, \forall_\varepsilon > 0$.

2.2.2 几乎处处收敛

现在考虑第二种精确陈述“正面朝上的比例接近 50%”的方法：只要持续抛掷硬币，那么对于任何能观察到的硬币抛掷序列，百分比 Y_n 均收敛到 50%．第一个证明这一结论的是雅可比·伯努利.



图 2-3 雅可比·伯努利（1655—1705）

事实上，在很多抛硬币的序列中，正面朝上的百分数不接近 50%. 比如，每次都出现“正面”的序列使得 $Y_n = 1$，所以 Y_n 不收敛于 50%. 类似地，序列 001001001001001… 使得 Y_n 收敛于 1/3 而非 1/2. 如果结合这些序列的存在以及上述的论断，我们得出的结论是，这些使得 Y_n 不收敛于 50%的序列的总概率为 0，也就是我们永远不会观察到它们的出现.

如果认真思考一下，就会发现这事实上是一个非常艰深的结论，因为需要排除的序列非常多，全部罗列基本上不可能. 但事实上，这个结论有一个十分巧妙的证明方式，我们下面就来看看证明是如何完成的. 首先，

$$P(|Y_n - 0.5| \geqslant \varepsilon) \leqslant E(\frac{|Y_n - 0.5|^4}{\varepsilon^4}), \forall n, \varepsilon > 0 .$$

与上面的例子相同，这个不等式是基于 $1\{|Y_n - 0.5| \geqslant \varepsilon\} \leqslant \frac{|Y_n - 0.5|^4}{\varepsilon^4}$ 并通过取期望得到的. 由上式可以得到：

$$E(|Y_n - 0.5|^4) = E(\frac{((X_0 - 0.5) + \cdots + (X_{n-1} - 0.5))^4}{n^4}) .$$

现在取 $Z_m = X_m - 0.5$，得到

$$E(((X_0 - 0.5) + \cdots + (X_{n-1} - 0.5))^4) = E((\sum_{m=0}^{n-1} Z_m)^4) = E(\sum_{a,b,c,d} Z_a Z_b Z_c Z_d) .$$

这里的求和是遍历所有 $a, b, c, d \in \{0, 1, \cdots, n-1\}$，包括 n 个 Z_a^4 项，$n(n-1)$ 个 $Z_a^2 Z_b^2$ 项（$a \neq b$），以及其中至少一个因子 Z_a 没有重复的其他项. 后面的这些项的期望值为 0. 这是因为所有其他的 b、c、d 项都与 a 不同，所以由独立变量的期望可以得到 $E(Z_a Z_b Z_c Z_d) = E(Z_a)E(Z_b Z_c Z_d) = 0$. 由此可得

$$E(\sum_{a,b,c,d} Z_a Z_b Z_c Z_d) = nE(Z_0^4) + n(n-1)E(Z_0^2 Z_1^2) = n\alpha + n(n-1)\beta .$$

其中 $\alpha = E(Z_0^4)$，而 $\beta = E(Z_0^2 Z_1^2)$. 将这一结果带入前面的表达式中，可以得到

$$P(|Y_n - 0.5| \geqslant \varepsilon) \leqslant \frac{n\alpha + n(n-1)\beta}{n^4 \varepsilon^4} \leqslant \frac{n^2(\alpha + \beta)}{n^4} = \frac{\alpha + \beta}{n^2} .$$

由此不等式得到①

$$\sum_{n \geqslant 1} P(|Y_n - 0.5| \geqslant \varepsilon) < \infty .$$

这个表达式表明，事件 $A_n := \{|Y_n - 0.5| \geqslant \varepsilon\}$ 的概率之和为有限数. 现在由附录 A.1 中的

① 回顾：$\sum_n \frac{1}{n^2} < \infty$.

博雷尔–坎泰利（Borel-Cantelli）引理得到

$$P(A_n, \text{i.o.}) = 0.$$

这个结果说明，ω 仅在有限个 A_n 中出现的概率为 1[①]. 所以，必然存在一个 $n(\omega)$使得 $n \geqslant n(\omega)$时 $\omega \notin A_n$. 也就是说，

$$|Y_n(\omega) - 0.5| \leqslant \varepsilon, \forall n \geqslant n(\omega).$$

由于上述的结论对于任意 $\varepsilon > 0$ 均成立，可以得到当 n 趋向无穷时，$Y_n(\omega) \to 0.5$.

实际上，如果 $Y_n(\omega)$不收敛于 50%，一定存在 $\varepsilon > 0$ 使得$|Y_n - 0.5| > \varepsilon$ 对无限多个 n 成立，但是我们已经看到情况并不是这样.

2.3 独立同分布随机变量的大数定律

上文抛掷硬币的结果可以推广到独立同分布随机变量$\{X_n, n \geqslant 0\}$上. 具体来说，要证明当 $n \to \infty$时，$Y_n := \dfrac{X_0 + \cdots + X_{n-1}}{n}$ 趋近于 $E(X_0)$. 如同抛掷硬币的例子一样，这里也有两种严格的数学方法来陈述这一结论.

2.3.1 弱大数定律

首先定义以下收敛模式.

定义 2.1 依概率收敛

令 $X_n, n \geqslant 0$ 和 X 为定义在同一个概率空间上的随机变量. 如果对于任意 $\varepsilon > 0$，当 $n \to \infty$ 时 $P(|X_n - X| \geqslant \varepsilon) \to 0$，则 X_n 依概率收敛到 X 并记作 $X_n \xrightarrow{p} X$. ◊

现在来看弱大数定律.

定理 2.2 弱大数定律

设$\{X_n, n \geqslant 0\}$是均值为 μ 的独立同分布随机变量序列，那么

$$Y_n = \frac{X_0 + \cdots + X_{n-1}}{n} \xrightarrow{p} \mu. \tag{2.4}$$

□

证明 假设 $E(X_n^2) < \infty$，其证明与抛掷硬币例子中的推导一样. 因此将其留给读者作为练习. 对于一般情况，请参见定理 13.18. ♣

---

① 可以通过令 $n(\omega) - 1$ 为满足 $\omega \in A_n$ 的最大 n 得到.

2.3.2　强大数定律

现在定义另一种收敛方式.

定义 2.3　几乎处处收敛

设$\{X_n, n\geqslant 0\}$和X为定义在同一个概率空间上的随机变量. 如果$P(\lim_{n\to\infty} X_n(\omega) = X(\omega)) = 1$，则称$X_n$几乎处处收敛于$X$，记作$X_n \to X, a.s.$.　◊

这一收敛意味着当$n\to\infty$时，实数序列$X_n(\omega)$以概率 1 收敛于实数$X(\omega)$.

同定理 2.2 中一样设$\{X_n, n\geqslant 0\}$，可以得到如下结论[①].

定理 2.4　强大数定律

设$\{X_n, n\geqslant 0\}$是一组均值为μ的独立同分布随机变量. 那么当$n\to\infty$时，$\frac{X_0+\cdots+X_{n-1}}{n}\to\mu$的概率为 1.　□

因此，样本均值$Y_n := \frac{X_0+\cdots+X_{n-1}}{n}$以概率 1 收敛于期望值（见图 2-4）.

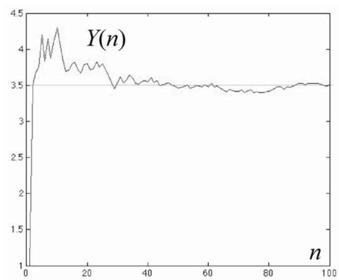


图 2-4　当投掷一枚均匀的骰子时，样本平均值收敛于 3.5

证明　假设$E(X_n^4) < \infty$，其证明与抛掷硬币例子中的推导相同. 因此将其留作练习. 一般情形下的强大数定律证明在定理 13.18 中给出.　♣

图 2-5 阐释了强大数定律和弱大数定律. 强大数定律指出，独立同分布随机变量的样本均值以 1 的概率收敛于均值. 弱大数定律告诉我们，当样本数量增加时，样本均值和期望值之间的差会变小.

---

① 几乎处处收敛可推出依概率收敛，所以强大数定律比弱大数定律更强. 见本章练习 5.

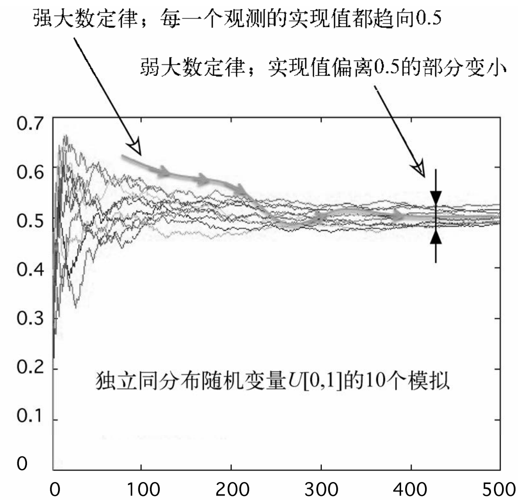


图 2-5 独立同分布随机变量 $U[0, 1]$的强大数定律和弱大数定律（另见彩插）

2.4 马尔可夫链的大数定律

对于一个有限不可约的马尔可夫链，它在某个状态停留的时间比例等于该状态的稳态分布值．例如一个在$\{0, 1\}$上的马尔可夫链 $X(n)$，设 $P(0, 1) = a = P(1, 0)$，其中 $a\in(0, 1)$．那么该马尔可夫链会有一半时间停留在状态 0．而图 1-1 中的马尔可夫链在状态 A 停留的时间比例为 12/39.

为了理解这一性质，我们来看看马尔可夫链回到某个状态 i 的时间间隔．图 2-6 绘出了一个由马尔可夫链产生的序列 $X(n)$．可以将这一序列拆分为从状态 i 开始再到状态 i 结束的循环．不难看到，由于马尔可夫链的性质，这些连续循环的时间长度 T_1, T_2, T_3, …是独立同分布的变量.

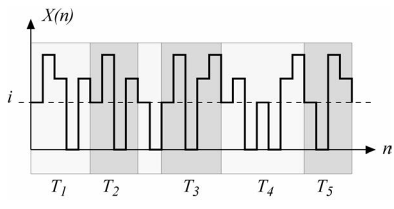


图 2-6 返回状态 i 的时长彼此是独立同分布的．大数定律揭示了在每个状态停留的平均时间比例最终收敛到一个常数

不难看到，这些随机时间长度的期望值都是有限的．事实上，对于固定的状态 i，任意的状态 j 都只需要有限步就能走到 i．定义 M_j 为所需要的最少步数．由马尔可夫

链的性质可知，存在一个非零的概率 p_j 使得只需要 M_j 步就能走到．记 $M = \max_j M_j$ 且 $p = \min_j p_j$．从任何一个状态开始的 M 步之后，马尔可夫链会以概率 p 走到状态 i．如果 M 步没有走到，那么重复刚刚的步骤．由此可知 $T_i \leqslant M\tau$，其中 τ 是以 p 为参数的几何变量．所以，$E(T_i) \leqslant ME(\tau) = M/p < \infty$．可以类似地得到 $E(T_i^4) \leqslant M^4E(\tau^4) < \infty$.

由以上的条件可以通过应用强大数定律得到：当 k 趋向无穷的时候，

$$\frac{T_1 + T_2 + \cdots + T_k}{k} \to E(T_1) \text{ 的概率为 } 1. \tag{2.5}$$

所以，马尔可夫链在状态 i 停留的长期时间比例以概率 1 收敛（详细推导见下文）：

$$\lim_{k\to\infty} \frac{k}{T_1 + T_2 + \cdots + T_k} = \frac{1}{E(T_1)}. \tag{2.6}$$

现在解释为什么式（2.6）等价于马尔可夫链在状态 i 上停留的时间比例收敛于 $1/E(T_1)$．设 $A(n)$为在时刻 n 前访问状态 i 的次数．首先证明 $A(n)/n$ 收敛于 $1/E(T_1)$．我们看到，对于 $T_1 + \cdots + T_k \leqslant n < T_1 + \cdots + T_{k+1}$，都有

$$\frac{k}{T_1 + \cdots + T_{k+1}} < \frac{A(n)}{n} = \frac{k}{n} \leqslant \frac{k}{T_1 + \cdots + T_k}.$$

如果假设当 $k \to \infty$时，$T_{k+1}/k \to 0$，由上述不等式可以得出

$$\frac{A(n)}{n} \to \frac{1}{E(T_1)}.$$

那么 T_{k+1}/k 是否趋向于 0？由上述的分析可得

$$P\left(\frac{T_{k+1}}{k} > \varepsilon\right) \leqslant P\left(\frac{M\tau}{k} > \varepsilon\right) \leqslant P(\tau > \alpha k) \leqslant (1-p)^{\alpha k},$$

其中 $\alpha = \varepsilon/M$.

因此，由博雷尔–坎泰利引理（见附录 A.1）可知，事件 $T_{k+1}/k > \varepsilon$ 在 k 趋向于无穷的时候仅出现有限次，也就证明了 $T_{k+1}/k \to 0$.

2.5 期望的收敛

设 X_n 和 X 是随机变量，并且 X_n 依概率收敛到 X. 那么这是否代表 $E(X_n) \to E(X)$？一般来说，答案是否定的．下面这个简单的例子就可以说明这一点．令[①] $Z =_D U[0, 1]$ 并定义 $X_n = n \cdot 1\{Z \leqslant 1/n\}$，其中 $n \geqslant 1$ 且 $X = 0$（见图 2-7）．可以观察到，若 $n > 1/Z$，则 $X_n = 0$．因此，n 足够大时有 $X_n = 0$，所以 $X_n \to 0$ 的概率为 1．X_n 等于 n 的概率为 $1/n$

① 记号“$=_D$”代表“和……有相同的分布”.

（当 $Z\leqslant 1/n$ 时），而在其他情况下 $X_n = 0$．因此 $E(X_n) = 1$ 对所有的 n 均成立．

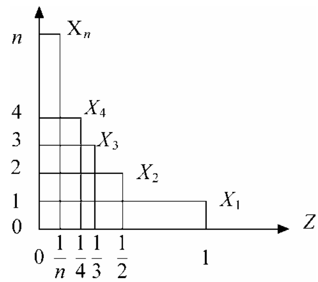


图 2-7　随机变量 X_n，使得 $X_n\to 0$ 且 $E(X_n)\to 1$

于是得到，X_n 依概率收敛到 X，且 $E(X_n) = 1 \not\to E(X) = 0$．

在一般情况下，取极限和求期望的顺序是不能调换的．也就是说，

$$\lim_{n\to\infty} E(X_n) \neq E(\lim_{n\to\infty} X_n) .$$

在前面的例子中，左边为 1，右边为 0．

下面以定理的形式给出两个期望的极限和极限的期望相等的充分条件．这两个条件将会在我们后续的分析中用到．不过在这里我们不予证明．

定理 2.5　控制收敛定理和单调收敛定理

设 X_n 和 X 满足 $X_n\to X$ 的概率为 1.

(a) 控制收敛定理（DCT）

若 $|X_n|\leqslant Y$ 且 $E(Y) < \infty$，则 $E(X_n)\to E(X)$.

(b) 单调收敛定理（MCT）

$$\text{若 } 0\leqslant X_n\leqslant X_{n+1},\ \forall\, n,\ \text{则 } E(X_n)\to E(X).$$

□

我们注意到，图 2-7 中的例子并不满足 MCT 和 DCT 的条件．

下面看 MCT 的应用，考虑一个事件序列 A_n．设 $X = \sum_{n=0}^{\infty} 1_{A_n}$ ．那么，由 MCT[①]可得

$$E(X) = E(\lim_{N\to\infty} \sum_{n=0}^{N} 1_{A_n}) = \lim_{N\to\infty} \sum_{n=0}^{N} E(1_{A_n}) = \sum_{n=0}^{\infty} P(A_n) .$$

这里特别值得注意的是[②]，

① 回顾：$\sum_{n=0}^{\infty} x_n := \lim_{N\to\infty} \sum_{n=0}^{N} x_n$.

② 若 $0\leqslant X$，$E(X) < \infty$，那么 $P(X < \infty) = 1$.

$$\text{如果}\sum_{n=0}^{\infty}P(A_n)<\infty\text{，则 }P(X<\infty)=1.$$

也就是说，ω 只属于有限个 A_n 的概率为 1．这也就是定理 A.1 中运用概率的连续性证明的博雷尔–坎泰利引理.

博雷尔–坎泰利的逆定理

这是博雷尔–坎泰利引理的部分逆定理.

引理 2.6　博雷尔–坎泰利的逆定理

设 A_n 为互相独立的事件，并有 $\Sigma_n P(A_n)=\infty$，则

$$P(A_n,\text{ i.o.})=1.$$

证明　设 $X_n=1\{A_n\}$.

$$\begin{aligned}P(X_n=0,\forall n\geqslant m)&=\lim_{n\to\infty}P(X_m=X_{m+1}=\cdots=X_n=0)\\&=\lim_{n\to\infty}(1-P(A_m))\cdots(1-P(A_n))\\&\leqslant\lim_{n\to\infty}\exp\{-P(A_m)\}\cdots\exp\{-P(A_n)\}\\&=\lim_{n\to\infty}\exp\{-P(A_m)-\cdots-P(A_n)\}=0.\end{aligned}$$

在第二步的不等式中，我们用到了 $1-x\leqslant\exp\{-x\}$，$\forall\, x\geqslant 0$.
所以得出结论，对于任意 m，可以以概率 1 找到一个 $n\geqslant m$，满足 $X_n=1$．也就是，存在无穷多个 $X_n=1$ 的概率为 1. ♣

下面的例子是这个逆定理的应用．考虑独立随机变量 X_n，这里 $X_n=B(1/n)$．由引理 2.6 可知存在无穷多个 $X_n=1$(因为概率和为无穷), 即 X_n 不收敛于 0 的概率为 1. 然而，对于任意 $\varepsilon>0$

$$\text{当 }n\to\infty\text{ 时，}P(|X_n-0|>\varepsilon)\leqslant P(X_n=1)=\frac{1}{n}\to 0.$$

因此，X_n 依概率收敛到 0，但不是几乎处处收敛.

2.6　大定理的证明

本节给出马尔可夫链重要结论的证明.

2.6.1　定理 1.2(a)的证明

设 m_j 为由状态 j 出发回到状态 j 的期望时间，即

$$m_j = E[T_j \mid X(0)=j]，其中\ T_j = \min\{n>0 \mid X(n)=j\}.$$

下面证明若马尔可夫链不可约，则 $\pi(j)=1/m_j$，$j=1, \cdots, N$ 是唯一的不变分布.

考虑一段时间 $n=1, \cdots, N$，$N \gg 1$. 根据定义，在这段时间里，马尔可夫链访问状态 j 的时间比例为 $1/m_j$. 在这些访问中，马尔可夫链访问状态 j 后紧接着访问状态 i 的比例为 $P(j,i)$. 所以，马尔可夫链在访问 j 后紧接着访问 i 的时间比例为 $P(j,i)\,(1/m_j)$. 通过对 j 求和，可以得到马尔可夫链访问 i 的时间比例为：

$$\sum_j \frac{1}{m_j} P(j,i) = \frac{1}{m_i}.$$

由此可得马尔可夫链的一个稳态分布为 $\pi_i = 1/m_i$，这正好是马尔可夫链访问状态 i 的时间比例.

如果还存在另一个稳态分布 $\phi(i)$，可以假设马尔可夫链从该分布开始. 于是有

$$\frac{1}{N}\sum_{n=0}^{N-1} 1\{X(n)=i\} \to \pi(i).$$

然而，假如对上式左右两边取期望，会发现左边等于 $\phi(i)$，因此 $\phi=\pi$. 由此可以看到 π 是马尔可夫链的唯一稳态分布[①].

2.6.2 定理 1.2(b)的证明

如果马尔可夫链是不可约的，但是具有周期性，则 π_n 有可能不收敛于稳态分布 π. 比如，如果一个马尔可夫链在 0、1 之间交替并且从 0 开始，那么当 n 为偶数时 $\pi_n=[1,0]$；当 n 为奇数时 $\pi_n=[0,1]$. 因此 π_n 不收敛于 $\pi=[0.5, 0.5]$.

如果马尔可夫链是非周期性的，那么 $\pi_n \to \pi$，而且收敛方式是几何收敛. 下面用图 2-8 所示的简单例子来解释这一结论. 考虑从状态 1 走回到状态 1 的步数 n：

$$\{n>0 \mid P^n(1,1)>0\} = \{3, 4, 6, 7, 8, 9, 10, \cdots\}.$$

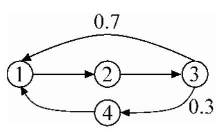


图 2-8 一个非周期性马尔可夫链

即当 $n \geqslant 6$ 时，$P^n(1,1)>0$. 由于 $P[X(2)=1 \mid X(0)=2]>0$，所以当 $n \geqslant 8$ 时，$P[X(n)=1 \mid X(0)=2]>0$. 事实上，若 $n \geqslant 8$，则 X 可以用 2 步从 2 到 1，然后用 $n-2$ 步从 1 回到 1. 对于所有其他的状态，都可以采用类似的证明. 由此，存在 $M>0$ 且 $p>0$ 使得

① 这里用了期望的性质，$E(1\{X(n)=i\}) = P(X(n)=i) = \phi(i)$.

$$P[X(M)=1|X(0)=i]\geqslant p,\quad i=1,2,3,4.$$

现在，考虑 2 个马尔可夫链：$\{X(n), n\geqslant 0\}$和$\{Y(n), n\geqslant 0\}$．其中 $X(0)$的分布为 π_0，而 $Y(0)$则服从稳态分布 π．让两个马尔可夫链独立演化．定义

$$\tau=\min\{n>0\mid X(n)=Y(n)\}.$$

由上述观察可得

$$P(X(M)=1 \text{ 和 } Y(M)=1)\geqslant p^2.$$

所以，$P(\tau>M)\leqslant 1-p^2$．如果 $\tau>M$，则 2 个马尔可夫链在时刻 M 还没有相遇．如前所述，可知 2 个马尔可夫链至少以概率 p^2 在接下来的 M 步中相遇．所以，

$$P(\tau>kM)\leqslant(1-p^2)^k.$$

现在修改 $X(n)$，使其在时刻 τ 与 $Y(n)$相遇后等于 $Y(n)$．注意这一耦合操作并不影响 $X(n)$，它仍然由转移阵 P 演变．因此，仍有 $P(X(n)=i)=\pi_n(i)$，$\pi_n=\pi_0P^n$．

由此可得

$$\sum_i\left|P(X(n)=i)-P(Y(n)=i)\right|\leqslant P(X(n)\neq Y(n))\leqslant P(\tau>n).$$

所以，

$$\sum_i|\pi_n(i)-\pi(i)|\leqslant P(\tau>n).$$

进一步得到

$$\text{若 } n>kM \text{，则} \sum_i\left|\pi_n(i)-\pi(i)\right|\leqslant(1-p^2)^k.$$

如果要将这一证明过程拓展到一般的非周期性马尔可夫链上，还需要证明，对于每个状态 i，存在整数 n_i，使得 $P^n(i,i)>0$ 对所有的 $n\geqslant n_i$ 均成立．我们会在下一节将这一结论作为引理 2.9 给予证明．

2.6.3 周期性

首先来看不可约马尔可夫链的返回时间的性质．

引理 2.7 考虑状态 i．设 $S:=\{n>0|P^n(i,i)>0\}$，并令 d 为 S 的最大公约数，那么集合 S 必然包含整数 n 和 n+d．

证明 这个证明很巧妙. 我们先用例子解释一下. 假设 $S=\{9,15,21,\cdots\}$并有 $d=g.c.d.(S)=3$．那么一定存在 $a,b\in S$ 使得 $g.c.d.\{a,b\}=3$．否则，S 的公约数不为 3．这里，取 $a=15$，$b=21$．考虑下述操作：

$$(a,b)=(15,21)\to(6,15)\to(6,9)\to(3,6)\to(3,3).$$

在每一步，将(x,y)（$x\leqslant y$）变成有序数对$(x,y-x)$．注意在每一步中，(x,y)的每一

项都是a和b的整数线性组合，比如$(6, 15) = (b-a, a)$. 那么$(6, 9) = (b-a, a-(b-a)) = (b-a, 2a-b)$，等等. 最终一定会得到(3, 3). 这些项会递减，直到得到 0. 假设得到(x, x)且$x\neq 3$，那么上一步有$(x, 2x)$，再上一步一定有$(x, 3x)$，一直回到(a, b)，所以a和b都是x的倍数. 但是$g.c.d.\{a, b\} = x$，出现了矛盾.

通过这种构建方式，有序对的每一项在每一步之后都是a, b的整数线性组合. 因此，$3 = ma+nb$，m和n为整数. 所以，

$$3 = m^+a+n^+b-m^-a-n^-b,$$

其中$m^+ = \max\{m, 0\}$，$m = m^+ - m^-$，$n = n^+ - n^-$. 现在设

$$N = m^-a+n^-b \text{ 且 } N+3 = m^+a+n^+b.$$

证明的最后一步利用S的以下性质：若$a, b\in S$，则$\alpha\, a + \beta\, b\in S$对于任意不全为 0 的整数$\alpha$和$\beta$成立. 这一性质可以通过$S$的定义得到. 所以，$N$和$N+3$都在$S$中.

对于一般的有公约数d的集合S，证明方法完全一样. ♣

这个结果可以用于证明马尔可夫链周期的存在.

引理 2.8 *对一个不可约的马尔可夫链，式（1.6）中定义的$d(i)$值对所有状态i都相同.*

证明 取$j\neq i$. 要证明该引理，只需证明$d(j)\leqslant d(i)$，因为由对称性也有$d(i)\leqslant d(j)$.

由马尔可夫链的不可约性可知，存在一个m使得$P^m(j, i) > 0$. 同样，也存在一个n使得$P^n(i, j) > 0$. 现在，由$d(i)$的定义和之前的引理可得，存在整数N使得$P^N(i, i) > 0$且$P^{N+d(i)}(i, i) > 0$. 但同时

$$P^{m+N+n}(j, j)>0 \text{ 和 } P^{m+N+d(i)+n}(j, j)>0.$$

这意味着整数$K := n+N+m$和$K+d(i)$都在集合$S := \{n > 0|P^n(j, j) > 0\}$中. 由此得到

$$d(j) := g.c.d.(S)\leqslant d(i).$$ ♣

可以通过证明下面的引理来证明马尔可夫链n步分布的收敛.

引理 2.9 *设X为一不可约的非周期性马尔可夫链. 令$S = \{n > 0|P^n(i, i) > 0\}$. 则存在$n_i$使得对于所有$n\geqslant n_i$都有$n\in S$.*

证明 首先，由于马尔可夫链的非周期性，我们有$d(i) = 1$. 从引理 2.7 可知存在整数N，使得$N, N+1\in S$. 下面证明

$$\text{任意的 } n > N^2 \text{ 都满足 } n\in S.$$

首先，对于$m > N-1$有

$$mN+0 = mN,$$

$$mN+1 = (m-1)N+(N+1),$$

$$mN+2 = (m-2)N+2(N+1),$$

$$\cdots,$$

$$mN+N-1=(m-N+1)N+(N-1)(N+1).$$

对于任意的 $n>N^2$，可以将其表示为

$$n=mN+k,$$

其中 $k\in\{0,1,\cdots,N-1\}$且 $m>N-1$. 因此, n 是 N 和 $N+1$ 的整数线性组合, 且 $N, N+1\in S$, 从而得到 $n\in S$. ♣

2.7　小结

- 样本空间
- 强大数定律和弱大数定律
- 控制收敛定理和单调收敛定理
- 由切比雪夫不等式推导出弱大数定律
- 由博雷尔–坎泰利引理与四次方期望约束推导出强大数定律
- 马尔可夫链独立同分布的返回时间的强大数定律
- 大定理的证明

重要方程与公式

强大数定律	$(X_1+\cdots+X_n)/n\to E(X_1)$的概率为 1	定理 2.4
控制收敛定理和单调收敛定理	$X_n\to X$，由两个收敛定理可得 $E(X_n)\to E(X)$	定理 2.5
切比雪夫不等式	$P(\|(X_1+\cdots+X_n)/n-\mu\|\geqslant\varepsilon)\leqslant \mathrm{var}(X_1)/\varepsilon^2$	式（2.2）
概率的收敛	$P(\|X_n-X\|\geqslant\varepsilon)\to 0$	定义 2.1
博雷尔–坎泰利引理	$\sum P(A_n)<\infty\Rightarrow P(A_n,\mathrm{i.o.})=0$	定理 A.1
马尔可夫链的强大数定律	$(1\{X_1=i\}+\cdots1\{X_n=i\})/n\to\pi(i)$的概率为 1	定理 1.2

2.8　参考资料

Chung 的 *Markov Chains with Stationary Transition Probabilities* 是一本非常好的马尔可夫链教材. 如果读者们希望找到进阶的概率理论教材，我会推荐 Billingsley 的 *Probability and Measure* , *Third Edition*.

2.9 练习

1. 考虑一个在$\{0, 1\}$中取值的马尔可夫链，解释为什么$\{0, 1\}$不是它的样本空间.

2. 考虑一个在$\{0, 1\}$中取值的马尔可夫链，其中 $P(0, 1) = a$，$P(1, 0) = b$．给出 2 个不同的样本空间及其概率分布，使得它们与该马尔可夫链相一致.

3. 绘出最小的周期性马尔可夫链. 证明马尔可夫链在各个状态中停留的时间比例收敛，但在时刻 n 处于某一状态的概率不收敛.

4. 对于问题 2 中的马尔可夫链，计算转移矩阵的特征值并用其计算马尔可夫链在时刻 n 的分布与稳态分布之间的距离范围.

5. 为什么由强大数定律可以推出弱大数定律？具体而言，设 X_n 和 X 为随机变量并且 X_n 几乎处处收敛到 X．证明 X_n 以概率收敛到 X.

提示：考虑一个 $\varepsilon > 0$ 并定义 $Z_n = 1\{|X_n-X|\geqslant\varepsilon\}$．运用控制收敛定理证明，若 X_n 几乎处处收敛到 X，则当 $n\to\infty$时 $E(Z_n) \to 0$.

6. 画一个 4 状态的不可约非周期性马尔可夫链. 考虑该链的 2 个独立版本：一个从状态 1 开始，另一个从状态 2 开始．说明为什么它们在有限的时间内会相遇.

7. 考虑图 1-1 中的马尔可夫链．用 Matlab 计算转移矩阵 P 的特征值．对于除 1 以外的特征值，设 λ 为其最大绝对值．再用 Matlab 计算

$$d(n) := \sum_i |\pi(i) - \pi_n(i)|,$$

其中，$\pi_0(A) = 1$．绘制 n 的函数 $d(n)$和 λ^n.

8. 抛掷一枚均匀的硬币．若为正面，可以得到随机数额为 X 的现金；若为反面，则可得到的数额为 Y．证明平均可得到$\frac{1}{2}E(X)+\frac{1}{2}E(Y)$的现金.

9. 你能找出几乎处处收敛到 0，但不以概率收敛到 0 的随机变量吗?

10. 令$\{X_n, n\geqslant 1\}$为独立同分布的随机变量，期望为 0 且方差为 σ^2. 证明：当 $n\to\infty$ 时，$X_n/n\to 0$ 的概率为 1.

提示：使用博雷尔–坎泰利引理.

11. 令 X_n 为建立在 $\mathcal{X}$ 上且稳态分布为 π 的不可约马尔可夫链，f: $\mathcal{X}\to\Re$ 为一函数．证明：

$$\text{当 } N\to\infty \text{ 时，}\frac{1}{N}\sum_{n=0}^{N-1} f(X_n) \to \sum_{i\in\mathcal{X}} \pi(i)f(i) \text{ 的概率为 1.}$$

第3章

多路复用—A

应用：通信链路共享，多径，缓冲器

主题：中心极限定理，置信区间，排队，随机协议

背景知识：附录 A.1～A.3，第 1 章

在通信网络中，允许多个用户共享通信资源是一个重要的概念.

例如，多个用户使用同一根同轴电缆；许多手机使用同一个基站；一个 Wi-Fi 热点为多个设备提供无线网络资源；在建筑或城市之间架设的高速链路可同时为很多用户传送数据.

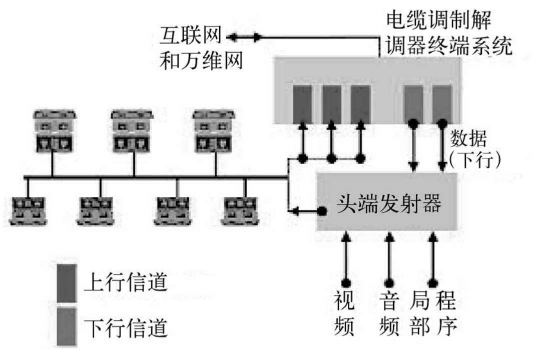


图 3-1　用于互联网接入的同轴线电缆（多用户共享）

这种多路复用共享机制通过在同一个物理媒介中同时传输属于不同用户的数据包来实现. 多路复用极大地降低了通信系统的成本. 本章将介绍多路复用的一些基本统计知识.



图 3-2　蜂窝基站的天线

在 3.1 节的开篇，我们将探讨当随机位用户共享同一个通信链路时，其中一位用户的信息传送速率. 这些计算对于网络的规划十分关键. 分析中用到的主要数学工具是中心极限定理. 为了共享一个通信链路，通信设备通常需要连接到交换机上. 例如，一幢建筑物中的台式机通常会被全部连接到一个交换机上，这些台式机的数据会被发送到一个共享的高速通信链路上. 3.5 节将会用到马尔可夫链为缓冲器建模并进行分析. 研究数据包通过交换机缓冲器时所经历的延时. 如果共享一个无线通信信道，通信设备之间则需要通过一个多址协议（multiple access protocol）来协调. 我们将在 3.6 节中用概率模型来分析这些协议.

3.1　链路共享

在互联网中，每一时刻都有很多数据流需要共享通信链路. 例如，可能会有 20 个用户同时通过同一个运营商的一根同轴电缆下载网页或视频资源.

在网络中，不同数据流的公平资源分配通常由传输控制协议（Transmission Control Protocol，TCP）来实现（至少理论上是这样）.

下面主要来考虑单一链路共享的情况. 如图 3-3 所示，这条链路的数据传输速率为 C bps（bits per second）. 若某个时刻有 v 位用户连接，那么每一位用户获得的数据传输速率为 C/v. 我们下面就来研究每位用户获得的数据传输速率. 这个问题的难点在于 v 是一个随机变量.

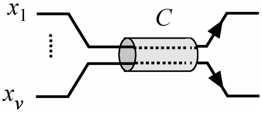


图 3-3　随机的 v 个连接共享一个数据传送速率为 C 的通信链路

首先考虑一个简单的模型. 假设一个通信链路有 $N \gg 1$ 位潜在的共享用户. 每位用户使用该链路的概率是 p 并且彼此独立. 那么，活跃用户（即使用链路的用户）的

数量 v 服从二项分布 $B(N, p)$（参见附录 A.2）.

图 3-4 给出了 $N = 100$ 与 $p = 0.1, 0.2, 0.5$ 时相应的概率质量函数. 对于 $N = 100$ 且 $p = 0.2$ 的情况，活跃用户数 v 为 $B(100, 0.2)$，也可以写作 $Binomial(100, 0.2)$. 平均来说，会有 $Np = 20$ 个活跃用户. 不过，也可能有多于 20 个活跃用户. 我们希望找到一个值 m，使得有多于 m 个活跃用户的概率是可以忽略的（小于 5%）. 这样就能以 95% 的概率保证每个活跃用户至少能获得 C / m 的数据传送速率.

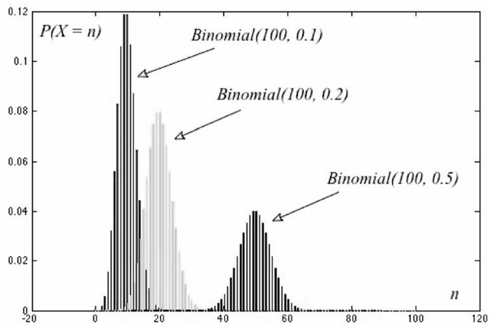


图 3-4　二项分布 $B(100, p)$ 对应的概率质量函数，其中 p 的取值分别为 0.1、0.2 和 0.5（另见彩插）

基于数值 m，就可以决定这个通信链路的带宽，以及对网络资源的规划. 直观估计，m 应当比流量的均值稍大. 我们可以看看图 3-5 所示的一个实际的分布（通过使用 Matlab 的 `disttool` 命令得到）. 可以看出：

$$P(v \leqslant 27) = 0.966 > 95\% \text{ 并且 } P(v \leqslant 26) = 0.944 < 95\%. \tag{3.1}$$

因此，使得 $P(v \geqslant m) \leqslant 5\%$ 的 m 最小值为 $m = 27$.

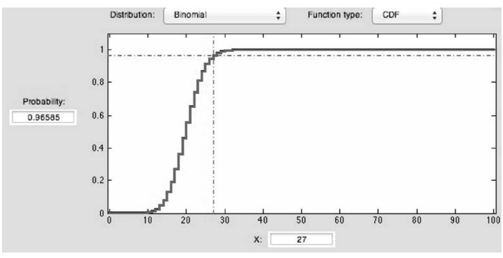


图 3-5　使用 MATLAB 的 `disttool` 命令可得式（3.1）的结论

下面来看可以用于近似二项分布的高斯分布. 我们稍后将根据这一近似对链路复用进行深入的探讨.

3.2 高斯随机变量与中心极限定理



图 3-6 卡尔·弗里德里希·高斯（1777—1855）

定义 3.1 高斯随机变量

(a) 如果随机变量 W 的概率密度函数(Probability Density Function，PDF) f_W 为

$$f_W(x) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x^2}{2}\}, x \in \Re .$$

则 W 被称为期望为 0、方差为 1 的高斯随机变量（或正态随机变量），记为 $W =_D \mathcal{N}(0,1)$．这时，也称 W 是一个**标准正态随机变量**或**标准高斯随机变量**.

(b) 若随机变量 X 满足

$$X = \mu + \sigma W ,$$

其中 $W =_D \mathcal{N}(0,1)$．则 X 被称为期望为 μ、方差为 σ^2 的高斯随机变量（或正态随机变量），记为 $X =_D \mathcal{N}(\mu,\sigma^2)$．相应地[①]，$X$ 的概率密度函数为

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\} .$$ ◇

图 3-7 为服从 $N(0,1)$ 分布的随机变量 W 的概率密度函数．特别注意

$$P(W > 1.65) \approx 5\%, P(W > 1.96) \approx 2.5\% \text{ 且 } P(W > 2.32) \approx 1\% . \quad (3.2)$$

① 参见附录 A.4.

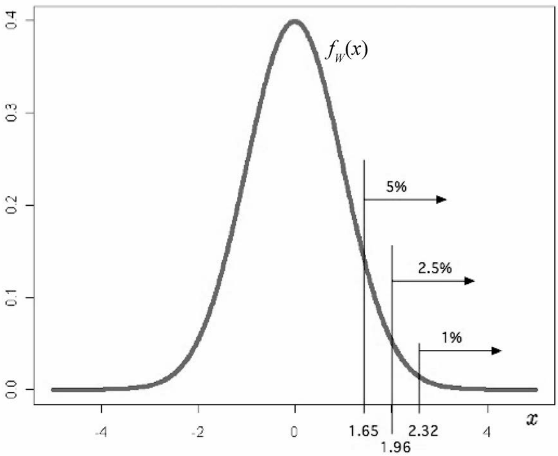


图 3-7　服从 $N(0,1)$ 分布的随机变量概率密度函数

中心极限定理指出，大量相互独立的随机变量之和近似于高斯分布．这个结论解释了为什么由许多电子随机扰动引起的热噪声服从高斯分布．许多其他的自然现象也都服从高斯分布，因为它们都是由许多相互独立的因素叠加而成的．

定理 3.2　中心极限定理

设 $\{X(n), n \geqslant 1\}$ 是独立同分布的随机变量，均值为 $E(X(n)) = \mu$，方差为 $\text{var}(X(n)) = \sigma^2$．那么，当 $n \to \infty$ 时，

$$\frac{X(1) + \cdots + X(n) - n\mu}{\sigma\sqrt{n}} \Rightarrow \mathcal{N}(0,1). \tag{3.3}$$

□

在式（3.3）中，符号"⇒"表示依分布收敛．具体来说，若 $\{Y(n), n \geqslant 1\}$ 是随机变量，那么 $Y(n) \Rightarrow \mathcal{N}(0,1)$ 表示

$$P(Y(n) \leqslant x) \to P(W \leqslant x), \forall x \in \Re .$$

其中 W 是服从 $\mathcal{N}(0,1)$ 的随机变量．我们会在下一节证明定理 3.2．

下面给出一般情况下依分布收敛的定义．

定义 3.3　依分布收敛

设 $\{X(n), n \geqslant 1\}$ 且 X 是随机变量．如果对所有使得 $P(X = x) = 0$ 的 x 都有

$$P(X(n) \leqslant x) \to P(X \leqslant x), \tag{3.4}$$

则称 $X(n)$ 依分布收敛于 X，记作 $X(n) \Rightarrow X$．　◊

举例来说，设 $X(n) = 3+1/n$（$n \geqslant 1$）且 $X = 3$．可以直观地看到 $X(n)$的分布收敛于 X 的分布．然而，

$$P(X(n) \leqslant 3) = 0 \nrightarrow P(X \leqslant 3) = 1 .$$

不过，

$$P(X(n) \leqslant x) \to P(X \leqslant x), \forall x \neq 3 .$$

这个例子解释了为什么定义中的式（3.4）只对使得 $P(X = x) = 0$ 的 x 要求 $P(X(n) \leqslant x)$ 收敛于 $P(X \leqslant x)$.

这里的收敛与概率论中的收敛以及处处收敛等概念有什么关系呢？首先，依分布收敛并不要求随机变量 $X(n)$与 X 在同一个概率空间中，因为定义 3.3 只与每一个随机变量的概率分布有关系．事实上，可以证明①

若 $X(n) \to X$ 以概率 1 收敛，则有 $X(n) \Rightarrow X$ （依分布收敛）.

因此，依分布收敛是最弱的一种收敛方式.

同时可以发现下面这个非常有用的结果：如果 $X(n) \Rightarrow X$ ，那么一定可以在同一个概率空间中构造随机变量 $Y(n)$和 Y，使得 $Y(n) =_D X(n)$， $Y =_D X$ ，且

$Y(n) \to Y$ 的概率为 1.

这个结论似乎有些神秘，但其实是很直观的．前面的章节提到，具有累积分布函数 $F(\cdot)$ 的随机变量可以通过选择随机变量 $Z =_D U[0,1]$ 并定义

$$X(Z) = \inf\{x \in \Re \mid F(x) \geqslant Z\}$$

来构造.

接下来可以验证 $P(X(Z) \leqslant a) = F(a)$．如果 $X(n) \Rightarrow X$ ，则当 $P(X = x) = 0$ 时有 $F_{X_n}(x) \to F_X(x)$．这意味着对所有 z

$X_n(z) = \inf\{x \in \Re \mid F_{X_n}(x) \geqslant z\} \to X(z) = \inf\{x \in \Re \mid F(x) \geqslant z\}$ 都成立.

二项分布与高斯分布

图 3-8 比较了二项分布与高斯分布.

为什么这两种分布如此相似呢？注意，如果 $X =_D B(N, p)$，则 X 可以写作

$$X = Y_1 + \cdots + Y_N ,$$

其中 Y_n 是参数为 p 独立同分布的伯努利随机变量．根据中心极限定理，有

$$\frac{X - Np}{\sqrt{N}} \approx \mathcal{N}(0, \sigma^2) .$$

① 见第 2 章练习 5 和本章练习 9.

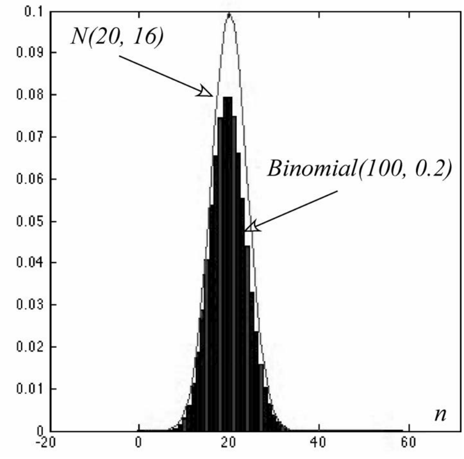


图 3-8　二项分布 $B(100,0.2)$ 与高斯分布 $N(20,16)$

其中 $\sigma^2=\mathrm{var}(Y_1)=E(Y_1^2)-(E(Y_1))^2=p(1-p)$ 为伯努利分布的方差．由此可以论证

$$B(N,p)\approx\mathcal{N}(Np,N\sigma^2)=\mathcal{N}(Np,Np(1-p))\,. \tag{3.5}$$

对于 $p=0.2$ 且 $N=100$ 的情况，有 $B(100,0.2)\approx\mathcal{N}(20,16)$，如图 3-8 所示．

3.3　多路复用与高斯分布

下面将二项分布的高斯分布近似应用到多址复用当中．回顾一下前面寻找最小值 m 的例子．我们采用以下的思路，由式（3.5）和式（3.2）可知：

(1)　$B(N,p)\approx\mathcal{N}(Np,Np(1-p)),\ (N\gg 1)$

(2)　$P(N(\mu,\sigma^2)\geqslant\mu+1.65\sigma)\approx 5\%$

于是，当 $N\gg 1$ 时，

$$P(B(N,p)\geqslant Np+1.65\sqrt{Np(1-p)})\approx 5\%\,.$$

因此，m 的最小值是

$$m=Np+1.65\sqrt{Np(1-p)}=20+1.65\sqrt{16}\approx 27\,.$$

从图 3-8 中也可以看出，当 $v=_D B(100,0.2)$ 时，v 几乎不可能大于 27.

3.4　置信区间

C.4

在上一节的计算中，也可以通过观察 $N\gg 1$ 时 $Y(N)$ 个活跃的用户来猜测 p 的取值．根据 3.3 节中(1)与(2)的思路，以及高斯分布关于均值的对称性，可以知道对于以

下两个事件：

$$A_1 = \{B(N,p) \geqslant Np + 1.65\sqrt{Np(1-p)}\}$$

以及

$$A_2 = \{B(N,p) \leqslant Np - 1.65\sqrt{Np(1-p)}\},$$

每个事件发生的概率都接近于 5%．对于 $Y(N) =_D B(N,p)/N$ ，A_1 与 A_2 相应地变成

$$A_1 = \{Y(N) \geqslant p + 1.65\sqrt{\frac{p(1-p)}{N}}\}$$

以及

$$A_2 = \{Y(N) \leqslant p - 1.65\sqrt{\frac{p(1-p)}{N}}\}.$$

因此，事件 $A_1 \cup A_2$ 的概率接近于 10%，这个事件的补事件概率则接近于 90%．所以

$$P(Y(N) - 1.65\sqrt{\frac{p(1-p)}{N}} \leqslant p \leqslant Y(N) + 1.65\sqrt{\frac{p(1-p)}{N}}) \approx 90\%.$$

尽管不知道 p 的取值，但是可以通过 $p(1-p) \leqslant 1/4$ 得到

$$P(Y(N) - 0.83\frac{1}{\sqrt{N}} \leqslant p \leqslant Y(N) + 0.83\frac{1}{\sqrt{N}}) \geqslant 90\%.$$

若 $N = 100$，则

$$P(Y(N) - 0.08 \leqslant p \leqslant Y(N) + 0.08) \geqslant 90\%.$$

例如，如果观察到 30%的用户是活跃的，则猜测 p 介于 0.22 和 0.38 之间的概率是 90%．也就是说，$[Y(N) - 0.08, Y(N) + 0.08]$ 是 p 的 90%置信区间．

图 3-7 表明，把 1.65 换成 2 就可以得到一个 5%的置信区间．所以

$$[Y(N) - \frac{1}{\sqrt{N}}, Y(N) + \frac{1}{\sqrt{N}}] \quad (3.6)$$

是 p 的 95%置信区间．

那么，N 取值多大才能很好地估计出 p 的值呢？例如希望有 95%的概率使 p 的取值在 $\pm 3\%$ 的范围内．由式（3.6），需要

$$\frac{1}{\sqrt{N}} = 3\%, \text{即 } N = 1089.$$

因此，$Y(1089)$是一个对 p 的估计值，其误差在 3%以内的概率是 95%．上面的这些结论其实正是大众意见调研问卷背后的理论依据．

在很多情况下，我们没办法像上述例子一样利用方差的界限．此时，可以将标准偏差替换为样本的标准偏差．也就是说，对于均值为 μ 的独立同分布的随机变量

$\{X(n), n \geqslant 1\}$，μ 的置信区间如下：

$$[\mu_n - 1.65\frac{\sigma_n}{\sqrt{n}}, \mu_n + 1.65\frac{\sigma_n}{\sqrt{n}}] = 90\%\text{ 置信区间}$$

$$[\mu_n - 2\frac{\sigma_n}{\sqrt{n}}, \mu_n + 2\frac{\sigma_n}{\sqrt{n}}] = 95\%\text{ 置信区间},$$

其中

$$\mu_n = \frac{X(1) + \cdots + X(n)}{n}$$

且

$$\sigma_n{}^2 = \frac{\sum_{m=1}^{n}(X(m) - \mu_n)^2}{n-1} = \frac{n}{n-1}\{\frac{\sum_{m=1}^{n}X(m)^2}{n} - \mu_n^2\}.$$

注意，这里的分母是 $n-1$．这是为什么呢？通常会认为样本方差是样本点偏离样本均值的平方值算术平均，即 $\sigma_n{}^2$ 第一个表达式中的分母 n．事实上，如果希望 $E(\sigma_n{}^2) = \sigma^2$，即得到没有偏差的估计量，应该除以 $n-1$ 而不是 n．对于很大的 n，显然这种区别是很小的．不过，还是应该先把为什么除以 $n-1$ 弄清楚．

为了便于表达，设 $E(X(n)) = 0$，令 $\sigma^2 = \text{var}(X(n)) = E(X(n)^2)$．注意到

$$\begin{aligned}
&n^2 E((X(1) - \mu_n)^2) \\
&= E((nX(1) - X(1) - X(2) - \cdots - X(n))^2) \\
&= E((n-1)^2 X(1)^2) + E(X(2)^2) + \cdots + E(X(n)^2) \\
&= (n-1)^2\sigma^2 + (n-1)\sigma^2 = n(n-1)\sigma^2.
\end{aligned}$$

对于第二个等式，注意交叉项 $E(X(i)X(j))$ 在 $i \neq j$ 的时候为 0（因为随机变量相互独立且均值都为 0）.

因此，

$$E((X(1) - \mu_n)^2) = \frac{n-1}{n}\sigma^2 \text{ 且} \sum_{m=1}^{n} E((X(m) - \mu_n)^2) = (n-1)\sigma^2.$$

所以，对 σ^2 的无偏差的估计量为

$$\sigma_n^2 := \frac{1}{n-1}\sum_{m=1}^{n} E((X(m) - \mu_n)^2).$$

3.5　缓冲器

互联网是一种包交换网络．主机（例如计算机）会把数据包发送给交换机．每台

交换机有多个输入和输出接口，如图 3-9 所示．交换机将接收到的数据包存在缓冲器中，然后根据它们的目的地通过合适的接口输出．因为数据包到达交换机的时间是任意的，所以有时候需要通过特定接口传输的数据包会集中在某段时间内到来．这种时候，交换机来不及处理所有的数据包．因此，数据包就会在缓冲器中累积起来．由于数据包需要在交换机的队列中等候，这就造成了延时．理解这种队列的延时对于理解交换机的原理十分重要．下面来研究这个系统的一个简单模型．

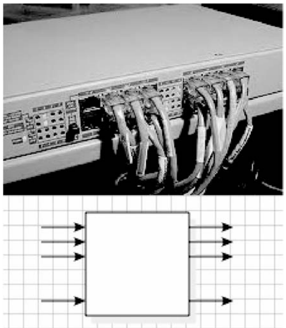


图 3-9 一台具有多个输入和输出接口的交换机

首先考虑从一个特定的输出端口传出数据包．假设交换机能为这个特定的输出端口保存 $N+1$ 个数据包．交换机在每个时刻可以传出一个数据包，但是也会有 $A(n)$个新数据包进入．这里的 $\{A(n), n\geqslant 1\}$ 为独立同分布的随机变量．

设 $X(n)$表示在时刻 n 存储的数据包数目．那么根据交换机的处理方式，有

$$X(n+1) = \max\{0, \min\{X(n)+A(n), N+1\}-1\}.$$

也就是说，$A(n)$个数据包到达交换机并与 $X(n)$个已经在缓冲器中的数据包汇合．当 $X(n)+A(n)$超过 $N+1$ 时，交换机需要丢弃无法储存的数据包．然后交换机会在缓冲器中的数据包中选择一个进行传输．在这些步骤之后，缓冲器中有 $X(n+1)$个数据包．

可以看到，序列 $\{X(n), n\geqslant 1\}$ 是一个定义在 $\mathcal{X} := \{0,1,\cdots,N\}$ 上的马尔可夫链．如果知道 $A(n)$取值为 k 的概率为 $p_k (k\geqslant 0)$，就可以写出 $X(n)$的转移概率．如果 $p_0 > 0$ 且 $p_2 > 0$，就可以很容易地验证这个马尔可夫链是不可约及非周期性的．然后可以计算马尔可夫链的稳态分布 π．同时可以知道稳态分布中的 $\pi(k)$ 表示马尔可夫链在 $X(n) = k$ 状态上停留的长期时间比例．

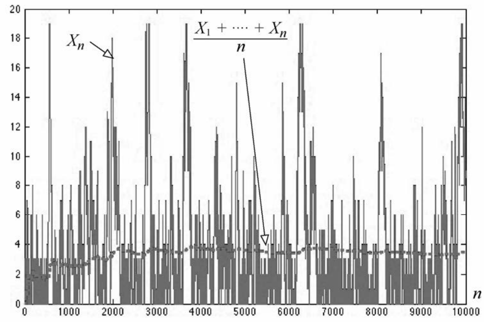


图 3-10　当 $A(n)$服从泊松分布 $P(0.9)$且 $N = 20$ 时，对队列长度 $A(n)$的仿真（也表示出了平均队列长度）（另见彩插）

通常来说，求解马尔可夫链的平衡方程并获得解析解是很困难的．下面来看一个可以求解的情况：

$$P(A_n = 2) = p \text{ 且 } P(A_n = 0) = 1 - p.$$

如果 $A_n = 2$，那么缓冲器占用量会增加 1；如果 $A_n = 0$，那么缓冲器占用量会减少 1. 这种情况下，$X(n)$的转移情况如图 3-11 所示.

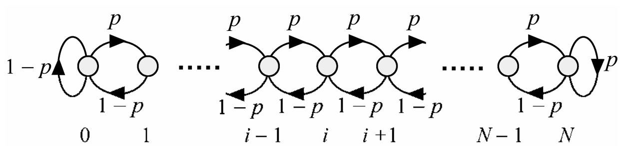


图 3-11　对于某一个输出端口，缓冲器占用量的转移概率

平衡方程组为

$$\begin{aligned}
&\pi(0) = (1-p)\pi(0) + (1-p)\pi(1) \\
&\pi(i) = p\pi(i-1) + (1-p)\pi(i+1), 1 \leqslant i \leqslant N-1 \\
&\pi(N) = p\pi(N-1) + p\pi(N).
\end{aligned}$$

可以验证以上问题的解为

$$\pi(i) = \pi(0)\rho^i, i = 0,1,\cdots,N \text{，其中 } \rho := \frac{p}{1-p}.$$

由于所有的概率之和为 1，有

$$\pi(0) = [\sum_{i=0}^{N} \rho^i]^{-1} = \frac{1-\rho}{1-\rho^{N+1}}.$$

还能计算出稳态分布下 $X(n)$的均值：

$$
\begin{aligned}
E(X(n)) &= \sum_{i=0}^{N} i\pi(i) = \pi(0)\sum_{i=0}^{N} i\rho^{i} \\
&= \rho[N\rho^{N+1} - (N+1)\rho^{N} + 1][(1-\rho)(1-\rho^{N+1})]^{-1} \\
&\approx \frac{\rho}{1-\rho} = \frac{p}{1-2p}.
\end{aligned}
$$

以上近似在 $N \gg 1$ 且 $\rho < 1$，即 $p < 1/2$ 时有效（由于缓冲器容量有限，近似的效果受缓冲器装满的概率影响）.

图 3-12 给出了在 $p = 0.45$ 且 $N = 20$ 时，对队列的仿真．它同时给出了 n 步之后队列的平均长度：随着 n 的增大，平均长度收敛到 $4.5 = p/(1-2p)$．注意这个队列几乎一直不是满的．因此可以认为缓冲器的容量相对较大．这也是可以在 $E(X(n))$表达式中让 $N \to \infty$ 的原因.

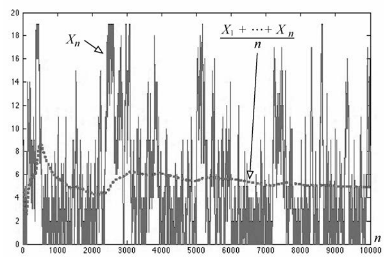


图 3-12　对 $p = 0.45$ 且 $N = 20$ 下队列的仿真（另见彩插）

那么数据包会在交换机中等待多长时间呢？在时刻 n，交换机中有 $X(n) = k$ 个数据包. 假如有 2 个数据包到达且 $k = 0$,那么第一个数据包在同一时刻 n 从交换机输出，第二个数据包在时刻 n+1 从交换机输出，于是平均延时就是(0+1)/2 = 0.5．如果 $k > 0$，那么第一个数据包在时刻 $n+k$ 从交换机输出，第二个数据包在时刻 $n+k+1$ 从交换机输出，平均延时为 k+0.5．因此，平均延时为 $X(n)$+0.5．同时，由于有 2 个数据包到达的概率不依赖于现在的队列长度，平均来说，到达的数据包发现队列中已经有 k 个数据包的概率等于稳态概率 $X(n) = k$．由此可以总结出，每个数据包的平均延时大约是 T：

$$
T = \frac{p}{1-2p} + 0.5.
$$

这时候，结合数据包的平均到达率是 $\lambda = 2p$ 可以发现

$$
E(X_n) = \lambda T .
$$

也就是说，队列中的平均积压量是到达率与平均延时的乘积. 这是利特尔定理（Little's Law）的一个具体例子，出自利特尔发表于 1961 年的文章“A Proof for the Queuing Formula: $L = \lambda W$ ”.

定理 3.4 利特尔定理

在一个排队系统中（弱假设下），

$$L = \lambda T \text{ ,}$$

其中 L 是系统中客户数目的平均值，λ 是客户的平均到达率，T 是客户在系统中花费的平均时间. □

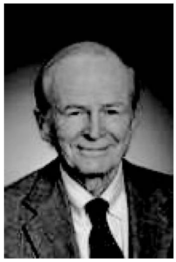


图 3-13 约翰 · 利特尔（1928— ）

可以这样直观地理解这个定理：考虑一个数据包在等待 T 个单位时间以后离开交换机的情况. 在它等待的时间里，平均有 λT 个数据包到达. 所以交换机中的平均积压量为 λT .

事实上，利特尔定理广泛地适用于各种系统，甚至包括不按数据包到达顺序进行处理的系统.

可以这么理解该结论. 假设每个数据包需要为它在交换机中停留的每一单位时间付 1 元钱. 如果数据包平均在交换机中停留 T 时间，那么每个数据包平均支付 T 元. 因为每单位时间有 λ 个数据包通过交换机，并且每一个数据包平均支付 T 元，所以交换机每单位时间收入 λT 元. 现在考虑另一种计算交换机收款率的方法. 在任意时刻，交换机中平均有 L 个数据包，因此每个时刻交换机的收款率为 L（因为数据包每等待一个单位时间就付一个单位的钱）. 由此可得 $L = \lambda T$.

3.6 多址访问

考虑如图 3-14 所示的场景：很多智能手机共享同一个 Wi-Fi 热点，并通过它传输数据包.

当多个智能手机同时传输的时候，它们会相互干扰并产生冲突. 因此，协调手机之间的传输十分关键. 下面来讨论一种能够协调这些传输并且成功率很高的简单方案. 我们用一个离散时间模型来对系统进行建模.

假设一共有 N 个设备. 在 n ($n \geqslant 1$)时刻,每个设备独立地以概率 p 进行数据传输. 这种传输方法最早是由诺曼·艾布拉姆森在 20 世纪 60 年代后期为 Aloha 网络提出的,被称为随机化多址访问(Randomized Multiple Access).



图 3-14 智能手机共享同一个 Wi-Fi 热点

在这个系统中,时刻 n 的传输数目 $X(n)$服从 $B(N, p)$(参见式(A.2)). 只有一个设备时的传输概率为

$$P(X(n)=1) = Np(1-p)^{N-1}.$$

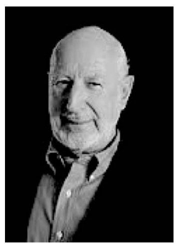


图 3-15 诺曼·艾布拉姆森(1932—)

以上概率在 $p = 1/N$ 取最大值. 这也就是系统支持的最大流量,记作 λ^*:

$$\lambda^* = (1-\frac{1}{N})^{N-1} \approx \frac{1}{e} \approx 0.36.$$

在上面的推导中,用到了如下结论:

$$(1-\frac{a}{N})^N \approx e^{-a}, \quad N \gg 1. \tag{3.7}$$

因此,这个方法的最大传输率约为 36%. 不过,这要求 $p = 1/N$,意味着每个设备都需要知道系统中有多少设备计划进行传输(即活跃用户数量). 在下一章里,我们将会讨论一种不需要知道这些信息的自适应方法.

3.7 小结

- 高斯随机变量 $N(\mu, \sigma^2)$

- ❑ 中心极限定理
- ❑ 置信区间
- ❑ 缓冲器：平均积压量、延时；利特尔定律
- ❑ 多址访问协议

重要方程与公式

$N(\mu,\sigma^2)$ 的定义	$f_X(x)=(2\pi\sigma^2)^{-1/2}\exp\{-(x-\mu)^2/(2\sigma^2)\}$	定义3.1
中心极限定理	$(X_1+\cdots+X_N-n\mu)/\sqrt{n}\Rightarrow\mathcal{N}(0,\sigma^2)$	定理3.2
95%置信区间	$(X_1+\cdots+X_n)/n\pm 2\sigma$	3.4节
利特尔定律	$L=\lambda W$	定理3.4
指数近似	$(1-a/n)^n\approx\exp\{-a\}$	式（3.7）

3.8　参考资料

对缓冲器的分析是排队论的一个简单应用．关于排队论在计算机以及通信系统中的模型请参考 Kleinrock 的 *Queueing Systems*（卷 1 和卷 2）．Kelly 的 *Reversibility and Stochastic Networks* 一书对排队网络的理论有很棒的解释，并已有电子书．Kelly 和 Yudovina 的 *Lecture Notes on Stochastic Networks* 探讨了最新的成果，是一份很精彩的课件．Srikant 和 Ying 的 *Communication Networks*: *An Optimization*, *Control*, *and Stochastic Networks Perspective* 解释了网络优化以及其他性能评估方面的问题．Walrand 的 *An Introduction to Queneing Networks* 中采用的方法也许比较难懂，但是很有帮助．

3.9　练习

1. 写一段 Matlab 代码来计算参与某项提案民意调查所需的人数，要求至少有 $1-\beta$ 的概率使得估计中赞成人数比例的误差在 α 之内．假设已知 $p\in[0.4,0.7]$，使用方差的上界进行计算．　C.4

2. 假设开展一项民意调查，估计支持候选人 Whatshisname 当选下一届总统的民众比例 p．其中调查了 N_1 个受过大学教育的人和 N_2 个没受过大学教育的人，支持 Whatshisname 先生当选的概率分别为 $B(p_1)$ 和 $B(p_2)$．在总人口当中，受过大学教育的人所占比例为 q．

   (a) 使用方差的上界来计算 p 的 95%置信区间．

   (b) 如果 $N_1+N_2=N$，应当怎样选取 N_1 和 N_2，才能够使得这个区间的长度最小？

3. 抛掷一枚均匀的硬币 10 000 次，正面朝上多于 5085 次的概率大约是：

- ❑ 15%
- ❑ 10%
- ❑ 5%
- ❑ 2.5%
- ❑ 1%

4. 用 Matlab 对缓冲器进行仿真，数据包到达的过程服从参数为 λ 的伯努利分布，为每个数据包提供的服务时间服从参数为 μ 的几何分布．画出仿真图并计算缓冲器的平均数据包积压量．

5. 考虑一个能够并行传输 M 个数据包的缓冲器．也就是说，当缓冲器内有 m 个数据包的时候，缓冲器能传输 $\min\{m, M\}$ 个数据包．并且每个数据包的传输在下一个时间节点结束的成功概率是 μ．在每一个时刻，有一个新数据包到达的概率是 λ．

C.3

(a) 这个马尔可夫链的转移概率是多少？

(b) 若希望系统稳定，即缓冲器中的数据包平均值为有限值，应当怎样为 λ、M 和 μ 取值？

(c) 用 Matlab 为这个系统仿真．

6. 为了估计一枚硬币正面向上的概率 p，将其抛掷 n 次，发现有 S_n 次正面向上．估计值为 $\beta = S_n/n$．通过选择样本大小 n 来保证

$$P(|S_n / n - p| \geqslant \varepsilon) \leqslant \delta .$$

(a) 根据切比雪夫不等式，n 的取值应当为多少（使用方差的上界进行计算）？

(b) 当 ε 变为 $\varepsilon / 2$ 时，n 的取值如何变化？

(c) 当 δ 变为 $\delta / 2$ 时，结果又会如何变化？

(d) 将这里 n 的取值与根据中心极限定理得出的取值作比较．

7. 设 $\{X_n, n \geqslant 1\}$ 独立同分布于 $U[0, 1]$，设 $Z_n = X_1 + \cdots + X_n$，那么 $P(Z_n > n)$ 是多大？如果用中心极限定理来推导，这个概率的估计值是多大？

8. 考虑一个缓冲器．假设每 2 秒钟有一个数据包到达，而传输一个数据包需要 1 秒钟（若缓冲器不为空则一直进行传输）．每个数据包在队列中的平均延时是多少？现在，假设每过 20 秒会有 10 个数据包到达，重新考虑以上问题．这个例子说明，队列的延时取决于数据到达过程的爆发性．

9. 证明：如果 $X(n)$ 依概率收敛到 X，那么 $X(n) \Rightarrow X$（即依分布收敛）．

提示：假设 $P(X=x)=0$. 为了证明 $P(X(n) \leqslant x) \to P(X \leqslant x)$，要注意若 $|X(n) - X| \leqslant \varepsilon$ 且 $X \leqslant x$，则 $X(n) \leqslant X + \varepsilon$．

第4章

多路复用—B

主题：特征方程，中心极限定理的证明，适应性的载波侦听多址传输

背景知识：附录 A，第 3 章

在第 3 章中，我们应用中心极限定理算出了可以安全共享电缆或连接线的用户数目．在这一章里，我们将证明这一定理．一个关键的数学工具是特征方程，它为研究独立随机变量的和提供了简便的方法．

4.1 特征方程

在证明中心极限定理之前，首先给出特征方程的解释．

定义 4.1 特征方程

随机变量 X 的特征方程定义如下：

$$\phi_X(u) = E(\mathrm{e}^{iuX}), u \in \Re,$$

其中 $i := \sqrt{-1}$． ◊

注意

$$\phi_X(u) = \int_{-\infty}^{\infty} \mathrm{e}^{iux} f_X(x)\mathrm{d}x,$$

这意味着 $\phi_X(u)$ 是 $f_X(x)$ 的傅里叶变换．由傅里叶变换的性质可知，由特征方程就可以唯一确定相应的概率密度函数．

下面的结论可以由特征方程推导而来，是一个非常有用的结论．

定理 4.2 $\mathcal{N}(0,1)$ 的特征函数

设 $X =_D \mathcal{N}(0,1)$，则

$$\phi_X(u) = e^{-\frac{u^2}{2}} \tag{4.1}$$

□

证明 由于

$$\phi_X(u) = \int_{-\infty}^{\infty} e^{iux} \frac{1}{2\sqrt{\pi}} e^{-\frac{x^2}{2}} dx,$$

因此，

$$\begin{aligned}\frac{d}{du}\phi_X(u) &= \int_{-\infty}^{\infty} ixe^{iux} \frac{1}{2\sqrt{\pi}} e^{-\frac{x^2}{2}} dx = -\int_{-\infty}^{\infty} ie^{iux} \frac{1}{2\sqrt{\pi}} de^{-\frac{x^2}{2}} \\ &= \int_{-\infty}^{\infty} i \frac{1}{2\sqrt{\pi}} e^{-\frac{x^2}{2}} de^{iux} = -u\int_{-\infty}^{\infty} e^{iux} \frac{1}{2\sqrt{\pi}} e^{-\frac{x^2}{2}} dx \\ &= -u\phi_X(u).\end{aligned}$$

这里第三个等式由分部积分法得到．因此，

$$\frac{d}{du}\log(\phi_X(u)) = -u = -\frac{d}{du}\frac{u^2}{2},$$

所以

$$\phi_X(u) = Ae^{-\frac{u^2}{2}}.$$

因为 $\phi_X(0) = E(e^{i0X}) = 1$，可知 $A = 1$，这就证明了式（4.1）. ♣

下面来证明中心极限定理.

4.2 中心极限定理的证明（概要）

为了研究相互独立的随机变量的和，要计算它们的特征函数．设

$$Y(n) = \frac{X(1) + \cdots + X(n) - n\mu}{\sigma\sqrt{n}}, n \geqslant 1,$$

我们有

$$\begin{aligned}\phi_{Y(n)}(u) &= E(e^{iuY(n)}) = E(\prod_{m=1}^{n} \exp\{\frac{iu(X(m)-\mu)}{\sigma\sqrt{n}}\}) \\ &= [E(\exp\{\frac{iu(X(1)-\mu)}{\sigma\sqrt{n}}\})]^n \\ &= [E(1 + \frac{iu(X(1)-\mu)}{\sigma\sqrt{n}} - \frac{u^2(X(1)-\mu)^2}{2\sigma^2 n} + o(1/n))]^n \\ &= [1 - u^2/(2n) + o(1/n)]^n \to \exp\{-u^2/2\}, n \to \infty.\end{aligned}$$

第三个等式成立是因为 $X(m)$是独立同分布的，而第四个等式成立是由于指数函数的泰勒展开：

$$e^{a} \approx 1 + a + \frac{1}{2}a^{2}.$$

因此，$Y(n)$的特征函数收敛到了 $\mathcal{N}(0, 1)$随机变量的特征函数．这意味着它们的逆变换，即 $Y(n)$的概率密度也将收敛到一个 $\mathcal{N}(0, 1)$随机变量的概率密度．最后一步推导可以严格证明，这里从略．♣

4.3 $\mathcal{N}(0,1)$的高阶矩

可以利用服从 $\mathcal{N}(0, 1)$分布的随机变量 X 的特征函数来计算它的高阶矩．第一步，应用指数函数的泰勒展开，

$$\phi_X(u) = E(e^{iuX}) = E(\sum_{n=0}^{\infty}\frac{1}{n!}(iuX)^n) = \sum_{n=0}^{\infty}\frac{1}{n!}(iu)^n E(X^n).$$

第二步，再次应用指数函数的泰勒展开，

$$\phi_X(u) = e^{-\frac{u^2}{2}} = \sum_{m=0}^{\infty}\frac{1}{m!}(-\frac{u^2}{2})^m.$$

第三步，比较上面两个式子中 u^{2m} 项的系数可以发现

$$\frac{1}{(2m)!}i^{2m}E(X^{2m}) = \frac{1}{m!}(-\frac{1}{2})^m.$$

所以[①]

$$E(X^{2m}) = \frac{(2m)!}{m!2^m}. \tag{4.2}$$

例如，

$$E(X^2) = \frac{2!}{1!2^1} = 1, E(X^4) = \frac{4!}{2!2^2} = 3 .$$

最后，我们注意到 u 的奇次幂项系数一定为 0，因此

$$E(X^{2m+1}) = 0, m = 0,1,2\cdots.$$

（这由 $f_X(x)$ 的对称性很容易得到．）我们还得到

$$\text{var}(X) = E(X^2) - E(X)^2 = 1.$$

① 这里应用到了 $i^{2m} = (-1)^m$.

4.4 两个独立同分布于$\mathcal{N}(0,1)$的随机变量平方和

设 X 和 Y 是两个独立同分布于 $\mathcal{N}(0, 1)$的随机变量．下面要证明

$$Z = X^2 + Y^2 =_D Exp(1/2).$$

设 θ 是向量(X, Y)的倾角，设 $R^2 = X^2 + Y^2$．如图 4-1 所示，我们得到

$$\mathrm{d}x\mathrm{d}y = r\mathrm{d}r\mathrm{d}\theta.$$

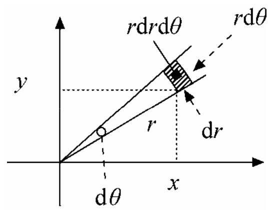


图 4-1 在极坐标变换 $x = r\cos(\theta)$ 和 $y = r\sin(\theta)$ 下，有 $\mathrm{d}x\mathrm{d}y = r\mathrm{d}r\mathrm{d}\theta$．也就是说，$[r, r+\mathrm{d}r]\times[\theta, \theta+\mathrm{d}\theta]$ 在(x, y)坐标平面上覆盖了 $r\mathrm{d}r\mathrm{d}\theta$ 的面积

因为 $E(Z) = E(X^2) + E(Y^2) = 2$，所以如果 Z 是服从指数分布的，它的参数一定是 1/2．下面来证明 Z 是指数分布的．我们有

$$\begin{aligned} f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y &= f_{X,Y}(x,y)r\mathrm{d}r\mathrm{d}\theta = \frac{1}{2\pi}\exp\{-\frac{x^2+y^2}{2}\}r\mathrm{d}r\mathrm{d}\theta \\ &= \frac{1}{2\pi}\exp\{-\frac{r^2}{2}\}r\mathrm{d}r\mathrm{d}\theta = \frac{\mathrm{d}\theta}{2\pi}\times\exp\{-\frac{r^2}{2}\}r\mathrm{d}r \\ &=: f_\theta(\theta)\mathrm{d}\theta\times f_R(r)\mathrm{d}r, \end{aligned}$$

其中

$$f_\theta(\theta) = \frac{1}{2\pi}1\{0<\theta<2\pi\}, \quad f_R(r) = r\exp\{-\frac{r^2}{2}\}1\{r\geqslant 0\}.$$

因此，(X, Y)的倾角 θ 和长度 $R = \sqrt{X^2+Y^2}$ 是相互独立的，分别具有上述的分布．对于 $V = R^2 =: g(R)$，根据定义，当 $v \geqslant 0$ 时，

$$f_V(v) = \frac{1}{|g'(R)|}f_R(r) = \frac{1}{2r}r\exp\{-\frac{r^2}{2}\} = \frac{1}{2}\exp\{-\frac{v}{2}\}.$$

这表明 θ 与 $V = X^2 + Y^2$ 相互独立．前者均匀分布在$[0, 2\pi]$，后者服从均值为 2 的指数分布．

4.5　特征函数的两个应用

我们已经用特征函数证明了中心极限定理. 下面介绍两个更有趣的应用.

4.5.1　泊松分布作为二项分布的近似

均值为 λ 、服从泊松分布的随机变量 X 可以用于近似 $n \to \infty$ 时的二项分布 $B(n, \lambda / n)$ 随机变量 X_n . 要证明这一点，注意到：

$$E(\exp\{iuX_n\}) = E(\exp\{iu(Z_n(1) + \cdots + Z_n(n))\}),$$

其中的随机变量 $\{Z_n(1), \cdots, Z_n(n)\}$ 是独立同分布的伯努利随机变量，均值为 λ / n . 因此，

$$E(\exp\{iuX_n\}) = [E(\exp\{iu(Z_n(1)\}))]^n = [1 + \frac{\lambda}{n}(\mathrm{e}^{iu} - 1)]^n.$$

上述第二个等式用到了如下结论：如果 $Z =_D B(p)$ ，则

$$E(\exp\{iuZ\}) = (1 - p)\mathrm{e}^0 + p\mathrm{e}^{iu} = 1 + p(\mathrm{e}^{iu} - 1).$$

同时，由于

$$P(X = m) = \frac{\lambda^m}{m!}\mathrm{e}^{-\lambda} \text{ 以及 } \mathrm{e}^a = \sum_{m=0}^{\infty} \frac{a^m}{m!},$$

我们有

$$E(\exp\{iuX\}) = \sum_{m=0}^{\infty} \frac{\lambda^m}{m!} \exp\{-\lambda\}\mathrm{e}^{ium} = \exp\{\lambda(\mathrm{e}^{iu} - 1)\}.$$

可以看到，

$$(1 + \frac{a}{n})^n \to \mathrm{e}^a, n \to \infty.$$

由此得出多项式分布的特征函数收敛到泊松分布的特征函数. 因此泊松分布可以很好地近似二项分布.

4.5.2　指数分布作为几何分布的近似

一个指数分布随机变量可以作为一系列服从几何分布随机变量的近似. 设 $X =_D Exp(\lambda)$ 且 $X_n = G(\lambda / n)$ ，则

$$\frac{1}{n} X_n \to X \text{ ，依概率收敛.}$$

为了说明这一点，首先由指数分布的性质得到

$$f_X(x) = \lambda e^{-\lambda x} 1\{x \geqslant 0\}.$$

同时，若 β 的实部为正，则

$$\int_0^{\infty} e^{-\beta x} = \frac{1}{\beta}.$$

因此，

$$E(e^{iuX}) = \int_0^{\infty} e^{iux} \lambda e^{-\lambda x} dx = \frac{\lambda}{\lambda - iu}.$$

另一方面，X_n 为几何分布，因此

$$P(X_n = m) = (1-p)^m p, m \geqslant 0.$$

由 $p = \dfrac{\lambda}{n}$ 可得

$$\begin{aligned}
E(\exp\{iu\frac{1}{n}X_n\}) &= \sum_{m=0}^{\infty}(1-p)^m p \exp\{ium/n\} \\
&= p\sum_{m=0}^{\infty}[(1-p)\exp\{iu/n\}]^m = \frac{p}{1-(1-p)\exp\{iu/n\}} \\
&= \frac{\lambda/n}{1-(1-\lambda/n)\exp\{iu/n\}} \\
&= \frac{\lambda}{n(1-(1-\lambda/n)(1+iu/n+o(1/n)))} \\
&= \frac{\lambda}{\lambda - iu + o(1/n)},
\end{aligned}$$

其中 $o(1/n) \to 0, n \to \infty$，从而证明了上述结论.

4.6 误差函数

在置信区间的计算中进行如下近似：

$$Q(x) := P(X > x)\text{，其中 } X =_D \mathcal{N}(0,1).$$

这里函数 $Q(x)$被称为*误差函数*. 它的值可以通过 Matlab 或者相应的手机应用进行计算. 下面给出的界限（见图 4-2）也很有用.

定理 4.3　误差函数的界

$$\frac{x}{1+x^2}\frac{1}{\sqrt{2\pi}}\exp\{-\frac{x^2}{2}\} \leqslant Q(x) \leqslant \frac{1}{x\sqrt{2\pi}}\exp\{-\frac{x^2}{2}\}, \forall x > 0. \qquad \square$$

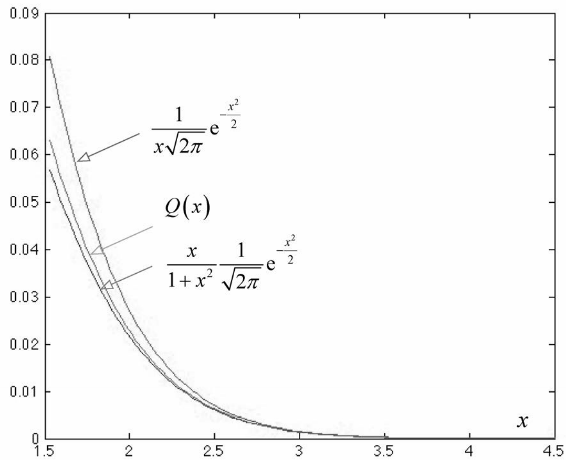


图 4-2　误差函数 $Q(x)$及其上下界（另见彩插）

证明　误差函数的上界可以如下导出．当 $x>0$ 时，

$$\begin{aligned}Q(x)&=\int_x^\infty f_X(y)\mathrm{d}y=\int_x^\infty\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{y^2}{2}}\mathrm{d}y=\frac{1}{\sqrt{2\pi}}\int_x^\infty\frac{y}{y}\mathrm{e}^{-\frac{y^2}{2}}\mathrm{d}y\\&\leqslant\frac{1}{\sqrt{2\pi}}\int_x^\infty\frac{y}{x}\mathrm{e}^{-\frac{y^2}{2}}\mathrm{d}y=\frac{1}{x\sqrt{2\pi}}\int_x^\infty y\mathrm{e}^{-\frac{y^2}{2}}\mathrm{d}y\\&=-\frac{1}{x\sqrt{2\pi}}\int_x^\infty\mathrm{d}\mathrm{e}^{-\frac{y^2}{2}}=\frac{1}{x\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}.\end{aligned}$$

对于下界，当 $x>0$ 时，

$$\begin{aligned}(1+\frac{1}{x^2})\int_x^\infty\mathrm{e}^{-\frac{y^2}{2}}\mathrm{d}y&\geqslant\int_x^\infty(1+\frac{1}{y^2})\mathrm{e}^{-\frac{y^2}{2}}\mathrm{d}y\\&=-\int_x^\infty\mathrm{d}(\frac{1}{y}\mathrm{e}^{-\frac{y^2}{2}})=\frac{1}{x}\mathrm{e}^{-\frac{x^2}{2}}.\end{aligned}$$

♣

4.7　自适应多址访问

在 3.6 节中，我们介绍了随机化多址访问机制．在该机制中，每一位用户在任意时间节点独立地以概率 $1/N$ 进行信号传输（N 为活跃用户个数），系统传输成功率大约是 $1/\mathrm{e}\approx 36\%$．不过，该机制要求用户知道系统中共有多少活跃用户．

为了使方案能够自动适应不同数量的活跃用户数，让传输设备在时间节点 n 按照下述方法调整传输的概率 $p(n)$：

$$p(n+1)=\begin{cases}p(n), & 若X(n)=1\\ ap(n), & 若X(n)>1\\ \min\{bp(n),1\}, & 若X(n)=0.\end{cases}$$

在这个方法中，a 和 b 为常数，满足 $a\in(0,1)$ 且 $b>1$．这里的思路是，如果某时刻没有用户进行传输，则下一时刻增大 $p(n)$；如果多个用户的传输产生冲突，则使 $p(n)$ 减小．这个方案是由布鲁斯・哈耶克（Bruce Hajek）和 Van Loon 于 1982 年提出的．

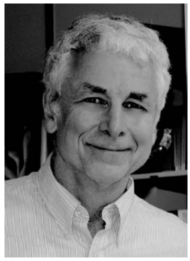


图 4-3　布鲁斯・哈耶克

图 4-4 展示了成功率 T_n 随时间的变化．这里，

$$T_n=\frac{1}{n}\sum_{m=0}^{n-1}1\{X(m)=1\}\ .$$

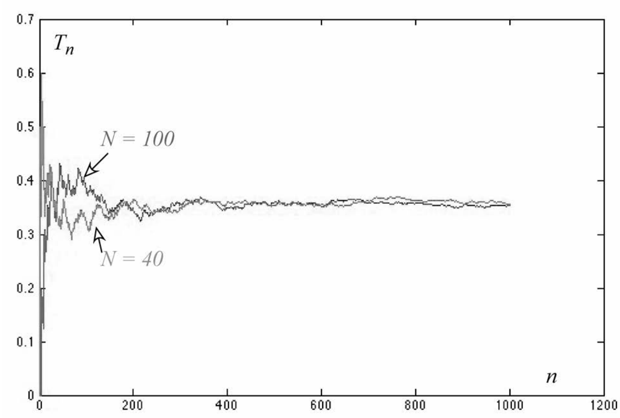


图 4-4　在自适应多址访问机制下，吞吐率关于时间的函数（另见彩插）

图中应用的参数是 $a=0.8$ 和 $b=1.2$．可以看到，吞吐率达到了 $N=40$ 和 $N=100$ 时的最优值．因此，这个方案可以自动根据活跃的设备数目进行调节．

4.8　小结

- 特征函数
- 中心极限定理的证明
- 高斯随机变量的高阶矩
- 高斯随机变量的平方和
- 泊松分布作为二项分布的近似
- 指数分布作为几何分布的近似
- 自适应多址访问协议

重要方程与公式

特征函数	$\phi_X(u) = E(\exp\{iuX\})$	定义4.1
$\mathcal{N}(0, 1)$分布	$\exp\{-u^2/2\}$	定理4.1
$\mathcal{N}(0, 1)$分布的高阶矩	$E(X^{2m}) = (2m)!/(m!2^m)$	式（4.2）
误差函数 $P(\mathcal{N}(0,1) > x)$	上下界	定理4.3

4.9　参考资料

中心极限定理是一个经典的概率结论，可以参阅 Bertsekas 和 Tsitsiklis 的 *Introduction to Probability*，Grimmett 和 Stirzaker 的 *Probability and Random Processes* 或 Billingsley 的 *Probability and Measure*，*Third Edition*.

4.10　练习

1. 设 X 是一个服从 $N(0, 1)$分布的随机变量．可以推知 $E(X^2) = 1$， $E(X^4) = 3$．

(a) 用切比雪夫不等式推导 $P(|X| > 2)$ 的一个上界．

(b) 用关于 X 的四阶矩的不等式来推导 $P(|X| > 2)$ 的上界. 可以找到更小的上界吗?

(c) 将上述结论与已知的 $N(0, 1)$随机变量知识作比较.

2. 为哈耶克的随机多址访问方案写一段 Matlab 仿真．假设有 20 个用户，在任意的时间节点，每个用户都有 $\lambda/20$ 的概率收到新信号．这些用户根据前文中陈述的规则更新其传输概率．请画出所有用户缓存的积压量随时间的变化.

3. 考虑这样一种多址访问方案：系统中的 N 个用户独立地以概率 p 传输用以预

定信道的数据包．这些数据包传输所需时间为 1 个时间单位．如果这些用以预定的数据包发生冲突或者没有用户传输预定数据包，用户会在下一个时刻再次尝试传输．如果预定成功，即只有 1 个用户传输用以预定的数据包，则该用户会在接下来的 K 个时间单位内独占信道并进行数据传输．当这次传输结束后，所有的用户重复以上过程．计算此机制下该信道能够传输的数据量．注意：这个方案叫作预约 ALOHA．

4. 设 X 是一个随机变量，均值为 0，方差为 1．证明 $E(X^4) \geqslant 1$．

提示：利用 $E((X^2-1)^2) \geqslant 0$．

5. 设 X 和 Y 是两个随机变量．证明：

$$(E(XY))^2 \leqslant E(X^2)E(Y^2).$$

这就是柯西–施瓦茨不等式.

提示：利用 $E((\lambda X - Y)^2) \geqslant 0$，令 $\lambda = E(XY)/E(X^2)$．

第5章

数字链路—A

应用：通过物理媒介传送比特数据

主题：MAP，MLE，假设检验

背景知识：附录 A，3.2 节

数字链路由发射机和接收机组成，通过物理媒介传送比特. 这些媒介可以是电缆、电话线、激光束、光纤、电磁波甚至是声波.



图 5-1　一个模拟通信系统

一个简单的系统[①]可由一根电话线构成. 传送比特 0 的时候，发射机在电话线末端将电压设为 −1 伏并持续 T 秒；而传送比特 1 的时候，则将电压设为 +1 伏并持续 T 秒. 接收机会测量线路末端的电压值. 如果接收机测得的电压为负，它会认为发射机送出比特 0；如果为正，则认为发射机送出比特 1. 由于传输过程中可能会受到干扰发生错误，接收机接收到的信号往往带有噪声并且信号的强度会有一些衰减. 因此，0 可能会被误认为 1，反之亦然. 为了减少这种错误的发生，人们设计了不同的编码方法.

本章探讨数字链路的工作原理和特性. 首先讨论贝叶斯准则和贝叶斯决策理论. 然后将这些想法应用到一个简单的通信链路模型上. 接下来，学习一种能使信号

① 这里省略了很多收发端进行时间同步的细节.

传输更快的编码方法. 最后, 本章对实际系统采用的调制和检测系统进行讨论, 其中包括异步数字用户线路和有线调制解调器.

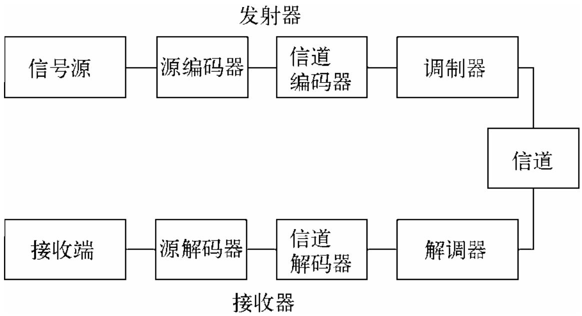


图 5-2 数字链路的组成

5.1 检测与贝叶斯准则

检测的定义是接收机接收到一个信号 S 并试图辨别发射机发送的内容. 对于这个问题, 我们首先通过一个一般性的模型进行探讨, 然后讨论具体的情况.

5.1.1 贝叶斯准则

考虑下面的基本模型: 有 N 个排他的起因能导致一个具体的症状出现. 这里的排他性指的是每次只有一个缘由发生. 每个起因 i 发生的先验概率为 p_i, 起因 i 发生以后观察到症状的概率为 q_i. 于是

$$p_i \geqslant 0, \sum_{i=1}^{N} p_i = 1, q_i \in [0,1].$$

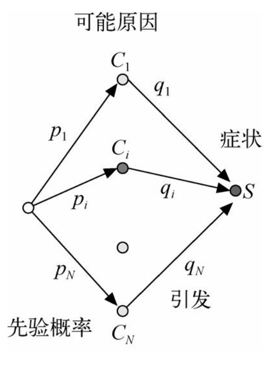


图 5-3 症状和可能导致这一症状的所有起因

贝叶斯准则可以用于计算由起因 i 引起症状的后验概率 π_i. 令 C_i 表示症状由起因 i 引起，并令 S 表示观察到症状. 那么，

$$\pi(i)=P[C_i\mid S]=\frac{P(C_i\text{和}S)}{P(S)}$$
$$=\frac{P(C_i\text{和}S)}{\sum_{j=1}^{N}P(C_j\text{和}S)}=\frac{P[S\mid C_i]P(C_i)}{\sum_{i=1}^{N}P[S\mid C_j]P(C_j)}$$
$$=\frac{p_iq_i}{\sum_{j=1}^{N}p_jq_j}.$$

这一结果非常重要，下面将其以定理的形式叙述出来.

定理 5.1　贝叶斯准则

设可观察到的症状 S 有 N 个排他的可能原因 C_i. 其中每个 C_i 的先验概率为 p_i，且 C_i 导致症状 S 发生的概率为 q_i. 那么，观测到的症状 S 由 C_i 引起的后验概率 π_i 为

$$\pi_i=\frac{p_iq_i}{\sum_{j=1}^{N}p_jq_j}. \tag{5.1}$$

□

这个准则十分简单，但是却是一个有关观测现象如何影响我们认知的经典例子.



图 5-4　托马斯 · 贝叶斯（1701—1761）

5.1.2　最大后验概率（MAP）与最大似然估计（MLE）

对于这个模型而言，可以看到最有可能导致症状出现的原因可以如下确定：

$$MAP=\arg\max_i \pi_i=\arg\max_i p_iq_i.$$

我们称之为给定症状原因的 MAP（Maximum A Posteriori，最大后验概率）估计.

这里的 argmax 含义如下：若 $h(\cdot)$是一个函数，那么 $\text{argmax}_x h(x)$则是使 $h(\cdot)$取最大值的变量 x. 所以，若 $x^*=\arg\max_x h(x)$，则 $h(x^*)\geqslant h(x)$ 对所有 x 均成立.

因此，MAP 代表的是最有可能的后验原因．如果所有的先验概率都相等，也就是对于所有 i 都有 $p_i = 1/N$，那么 MAP 代表 q_i 最大的原因．通常来说，最大化 q_i 的估计叫作对给定症状原因的 MLE（Maximum Likelihood Estimate，最大似然估计）．也就是，

$$MLE = \arg\max_i q_i \,.$$

MLE 是最可能导致症状发生的原因．

MAP 与 MLE 的一般定义如下．

定义 5.2　MAP 与 MLE

令(X, Y)为离散随机变量，那么

$$MAP[X \mid Y = y] = \arg\max_x P[X = x \mid Y = y]$$
$$MLE[X \mid Y = y] = \arg\max_x P[Y = y \mid X = x] \,.$$

后面会看到，这些定义可以很自然地延伸到连续情况．

◊

5.1.3　二元对称信道

现在将 MLE 和 MAP 的概念应用到简化的通信链路模型中．图 5-5 给出了这个通信模型，叫作二元对称信道（BSC）．

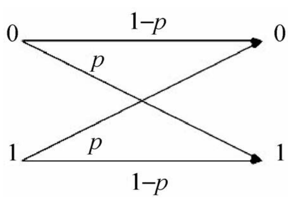


图 5-5　二元对称信道

在这个模型中，发射机发送信号 0 或 1，而接收机以 $1-p$ 的概率接收发送的比特，以 p 的概率接收相反的比特，即信道出错的概率为 p．每个比特出现的错误是独立同分布的．

如果 $p = 0$ 或 $p = 1$，那么可以准确地还原每一个接收到的比特．如果 $p = 0.5$，则信道的输出独立于输入．这种情况下，没有任何有用的信息通过信道．那么，在其他情况下会发生什么呢?

记 $X \in \{0, 1\}$为信道的输入，$Y \in \{0, 1\}$为其输出．假设观察到 $Y = 1$ 且已知 $P(X = 1) = \alpha$ 和 $P(X = 0) = 1-\alpha$．我们会得到图 5-6 所示的结果．

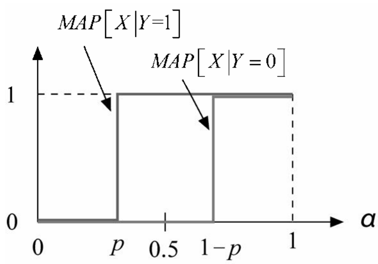


图 5-6　二元对称信道的最大后验估计．$\alpha = P(X=1)$，p 是信道出错的概率

定理 5.3　二元对称信道的 MPA 和 MLE

对于 $p < 0.5$ 的二元对称信道，

$$MAP[X|Y=0] = 1\{\alpha > 1-p\}，MAP[X|Y=1] = 1\{\alpha > p\}，$$

$$MLE[X|Y] = Y. \qquad \square$$

证明　运用贝叶斯准则的术语，事件 $Y=1$ 是症状，先验概率为

$$p_0 = 1-\alpha \text{ 和 } p_1 = \alpha.$$

同时，条件概率是

$$q_0 = P[Y=1|X=0] = p \text{ 和 } q_1 = P[Y=1|X=1] = 1-p.$$

因此，$MAP[X|Y=1] = \arg\max_{i\in\{0,1\}} p_i q_i$.

所以，

$$MAP[X|Y=1] = \begin{cases} 1, & 若 p_1q_1 = \alpha(1-p) > p_0q_0 = (1-\alpha)p \\ 0, & 其他情况. \end{cases}$$

由此可得，$MAP[X|Y=1] = 1\{\alpha > p\}$．也就是说，当 $Y=1$ 时，如果 $X=1$ 的先验概率大于信道出错的概率，就可以猜测 $X=1$.

并且

$$MLE[X|Y=1] = \arg\max_{i\in\{0,1\}} q_i.$$

在这种情况下，由于 $p<0.5$，总有 $MLE[X|Y=1]=1$．这是因为 $Y=1$ 在 $X=1$ 时比在 $X=0$ 时更容易发生．所以，最大似然估计忽视先验概率，当 $Y=1$ 时总是猜测 $X=1$，即便先验概率 $P(X=1)=\alpha$ 可能很小.

类似地，可以得到

$$MAP[X|Y=0] = \arg\max_{i\in\{0,1\}} p_i(1-q_i).$$

所以，

$$MAP[X \mid Y=0] = \begin{cases} 1, \text{若} p_i(1-q_1) = \alpha p > p_0(1-q_0) = (1-\alpha)(1-p) \\ 0, \text{其他情况} \end{cases}$$

由此可得 $MAP[X|Y=0] = 1\{\alpha > 1-p\}$．所以，当观察到 $Y=0$ 时，如果 $X=1$ 的先验概率比信道正确的概率要大，就可以猜测 $X=1$．

并且 $MLE[X|Y=0]=0$，因为 $p<0.5$． ♣

5.2 霍夫曼编码

编码可以提高数字链路的性能．在本节中，我们来探讨霍夫曼编码．

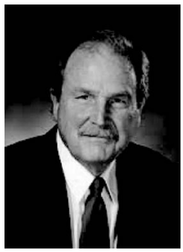


图 5-7　大卫・霍夫曼（1925—1999）

假设想要通过数字链路传送包含符号 *A*、*B*、*C*、*D* 的字符串．那么最简单的方法是将这些符号分别编码为 00、01、10 和 11．这样每个符号需要传送 2 比特．如果信道不存在错误，那么当接收机接到数据 0100110001，就能将信号解码并得出发送的字符串为 *BADAB*．

现在，假设字符串中符号出现的频率为(*A*, 55%)、(*B*, 30%)、(*C*, 10%)和(*D*, 5%)．因此，*A* 的出现占总时间的 55%，其他符号以此类推．在此情况下，一种好的编码机制应该使用比 *D* 更少的比特来对 *A* 进行编码．

对这个例子而言，霍夫曼编码[Huffman，1952]的编码方式如下：

$$A=0, B=10, C=110, D=111.$$

采用霍夫曼编码，每个符号平均所需的比特数为

$$1\times55\%+2\times30\%+3\times10\%+3\times5\% = 1.6.$$

因此，可以节省 20%的比特，系统也会快 25%．（计算方法的功劳！）注意，在信道没有错误时，这种编码使得接收机从接收到的比特中无误地恢复发送机传输的字符串．比如，如果接收机收到 110100111，符号就是 *CBAD*．不存在混淆的情况．

下面来看为什么解码的时候不会出现混淆．可以将比特串绘制成一棵树中到达叶子节点的路径，如图 5-8 所示．从第一个接收到的比特开始，沿树向下走直到到达一

个叶子，然后解码一个字符．然后再对剩下的比特重复上述步骤，直到比特串解码完成．在我们的例子中，比特串为 110100111．从树顶端开始，沿树枝 110 到达叶子 C，再从头开始沿树枝 10 到达叶子 B，以此类推．这种能够一次性无误解码的编码统称为无前缀编码．

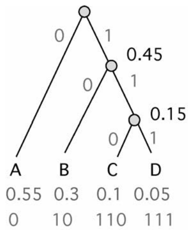


图 5-8　霍夫曼编码

无前缀编码的构建过程很简单．如图 5-8 所示，首先将出现频率最小的 2 个符号 C 和 D 结合为新节点 CD．新节点的 2 个分支为 0 和 1．然后将组合 CD 的权赋为这 2 个符号的频率之和 0.15．然后重复同样的步骤，结合节点 CD 和 B，并将组合 BCD 的权赋为 0.3+0.15 = 0.45．最后，结合 A 和 BCD．如此构建得到的树与编码相对应．

下面来关注霍夫曼编码的一个性质．

定理 5.4　霍夫曼编码的最优性

在所有无前缀编码中，霍夫曼编码使得每个字符所需的平均比特数最少．

□

证明　见第 6 章．♣

值得一提的是，其他编码可能达到更小的平均比特长度，但它们不是对逐个字符进行编码，采用的方式更为复杂．其中一种编码观察到在字符串长度 $n \gg 1$ 的时候，只有 2^{nH} 种可能的字符串．这里，

$$H = -\sum_X x\log_2(x)\,.$$

在这个表达式中，x 是符号 X 的频率，并对所有符号进行求和．这里的 H 是字符分布的熵．所以，可以通过列举所有的字符串并分配 nH 个比特来辨别它们，因而对于 n 个字符的编码只需要 nH 个比特，即每个字符 H 比特．

在我们的例子中，

$$H = -0.55\log_2(0.55) -0.3\log_2(0.3) -0.1\log_2(0.1) -0.05\log_2(0.05) = 1.54.$$

尽管在本例中最优编码相比于霍夫曼编码的优势并不明显，但是可以很容易地找到差别很大的例子．比如，假设只有两个字符 A、B，其频率分别是 p 和 $1-p$，$p\in(0,\ 1)$．霍夫曼编码中每个字符需要 1 比特，而基于长字符串的编码中每个字符需要$-p\log_2(p) -(1-p)\log_2(1-p)$

比特．当 $p=0.1$ 时，上式为 0.47，这比霍夫曼编码所需比特的一半还少．

我们会在 13.6 节中讨论基于长字符串的编码．

5.3 高斯信道

C.5

在前面的几节中，我们考虑了一个简化的信道模型：二元对称信道．本节考虑的信道模型则更能体现实际系统中噪声的物理特性．在这个模型中，发射机传送比特 $X\in\{0,1\}$，接收机接收到 Y，其中

$$Y=X+Z.$$

在这个等式中，$Z={}_D\mathcal{N}(0,\sigma^2)$与 X 无关．我们称这个信道模型为加性高斯白噪声信道．图 5-9 展示了当 $X=0$ 和 $X=1$ 时 Y 的分布．实际上，当 $X=x$ 时，$Y={}_D\mathcal{N}(x,\sigma^2)$.

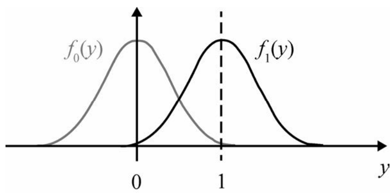


图 5-9 当 $X=0$ 时，Y 的概率密度函数是 f_0；$X=1$ 时，Y 的概率密度函数是 f_1

假设 $P(X=1)=p_1=\alpha$ 且 $P(X=0)=p_0=1-\alpha$．那么当接收机收到 Y，它要怎么判别 $X=0$ 还是 $X=1$？

在这个例子中，由于 Y 是一个连续随机变量，对于所有 y，$P[Y=y|X=0]=0$．因此，需要在应用贝叶斯准则时进行一些改变．首先假设每次测量 Y 的时候都观测到 $Y\in(y,y+\varepsilon)$，其中 $0<\varepsilon\ll 1$．在这种情况下，观测到的症状是 $Y\in(y,y+\varepsilon)$，并且发生的概率大于 0．事实上，

$$q_0=P[Y\in(y,y+\varepsilon)\mid X=0]\approx f_0(y)\varepsilon.$$

该式的约等号由 $X=0$ 时密度为 $f_0(y)$推导而来．同样，

$$q_1=P[Y\in(y,y+\varepsilon)\mid X=1]\approx f_1(y)\varepsilon.$$

因此，

$$MAP[X\mid Y\in(y,y+\varepsilon)]=\arg\max_{i\in\{0,1\}} p_i f_i(y)\varepsilon.$$

由于这个结果与 ε 无关，可以得到

$$MAP[X\mid Y=y]=\arg\max_{i\in\{0,1\}} p_i f_i(y).$$

同理，

$$MLE[X \mid Y=y] = \arg\max_{i\in\{0,1\}} f_i(y).$$

我们也可以验证

$$MAP[X \mid Y=y] = 1\{y \geqslant \frac{1}{2} + \sigma^2 \log(\frac{p_0}{p_1})\}. \quad (5.2)$$

产生错误的概率为：

$$P(\mathcal{N}(0,\sigma^2) \geqslant \frac{1}{2} + \sigma^2 \log(\frac{p_0}{p_1}))p_0 + P(\mathcal{N}(1,\sigma^2) \leqslant \frac{1}{2} + \sigma^2 \log(\frac{p_0}{p_1}))p_1.$$

并且，

$$MLE[X|Y=y] = 1\{y \geqslant 0.5\}.$$

在这个系统下，最大似然估计的错误概率与参数为 $p = p(\sigma^2) := P(\mathcal{N}(0,\sigma^2) > 0.5)$ 的二元对称信道的出错概率相同.

二元相移键控（BPSK）

前面的简单模型很好地对应了采用了 BPSK（Binary Phase Shift Keying，二元相移键控）传输机制的实际系统. 在这种系统中，发射机每次发送一个持续 T 秒的正弦波形，并根据发送的比特是 0 还是 1 决定其相位；而不是通过传送持续 T 秒的电平以代表 0 和 1 比特. 具体而言，如果传送的比特为 0，则发射机发送信号

$$\boldsymbol{s}_0 = \{s_0(t) = A\sin(2\pi ft), t \in [0, T]\}.$$

这里，T 是周期的倍数，且满足 $fT = k$（k 为整数）. 为传送比特 1，发射机发送信号 $\boldsymbol{s}_1 = -\boldsymbol{s}_0$.

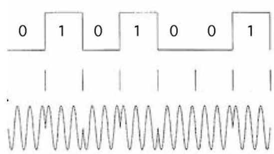


图 5-10　BPSK 下发射机发送的信号

为什么要这么麻烦呢？这么做的一个好处是，可以根据实际物理系统的需要选择合适的频段. BPSK 下的信号是在频率 f 附近的正弦波形，设计者可以选择适合传输媒介的频率. 比如，如果传输是无线的，就可以选择适合天线的辐射和接收的频率 f. 这里，发送的电磁波波长，也就是光速除以 f，需要和天线的物理尺寸在同一数量级. 比如，1GHz 与 1 英尺的波长相对应. 该频段的正弦波适合手机天线的发送和接收.

发射机通过传送信号 $\boldsymbol{s}_i$ 来发送比特 $i(i = 0, 1)$. 接收机则试图检测接收到的信号到

底是 $\boldsymbol{s}_0$ 还是 $\boldsymbol{s}_1 = -\boldsymbol{s}_0$. 为了做到这点，它将接收到的信号乘以在频率 f 处的正弦波形，然后计算乘积的均值. 如果接收机收到信号 $\boldsymbol{r} = \{r_t, 0 \leqslant t \leqslant T\}$，需要计算

$$\frac{1}{T}\int_0^T r_t \sin(2\pi f t)\,\mathrm{d}t.$$

这里可以验证，若 $\boldsymbol{r} = \boldsymbol{s}_0$，则结果为 $A/2$；若 $\boldsymbol{r} = \boldsymbol{s}_1$，则结果为 $-A/2$. 因此，如果均值为正，则接收机判定收到的信号为 0，否则为 1.

很多时候，当发射机发送 $\boldsymbol{s}_i$ 时，接收机接到的并不是 $\boldsymbol{s}_i$，而是具有一定衰减且带有噪声的信号. 因此，进行上述计算后，接收机得到 $B+Z$ 或 $-B+Z$. B 是取决于衰减的常量，而 Z 是服从 $\mathcal{N}(0, \sigma^2)$ 的随机变量，σ^2 代表噪声的能量. 于是，信号检测问题等同于前面讨论过的检测高斯随机变量的均值.

5.4 多维高斯信道

使用 BPSK 时，发射机有两种选择：$\boldsymbol{s}_0$ 和 $\boldsymbol{s}_1$. 所以，发射机在 T 秒内发送 1 比特. 为了增加传送速率，通信工程师发明了一种更有效率的传输机制：QAM（Quadrature Amplitude Modulation，正交幅度调制）. 采用这种传输方法，发射机可以在每 T 秒内发送 k 比特. 这一机制可根据不同的 k 值进行调整. 当 $k = 1$ 时，这种方法和 BPSK 完全一致；当 $k > 1$ 时，QAM 会采用 2^k 个不同的信号，每一个都形如

$$a\cos(2\pi f t) + b\sin(2\pi f t).$$

系数对 (a, b) 确定了信号的形式并且对应一个给定的 k 个比特的字符串. 这些系数组成一个如图 5-11 所示的星座图. 图 5-11 对应 16 进制正交幅度调制，也就是 $k = 4$ 的情况.

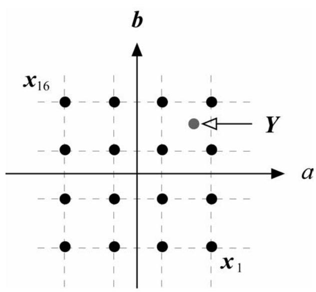


图 5-11　16QAM 星座图

接收机收到信号时，用 $2\cos(2\pi f t)$ 与之相乘并计算 T 秒内的平均值. 如果没有衰减和噪声，这个均值等于系数 a. 接收机也会用 $2\sin(2\pi f t)$ 与信号相乘，并计算 T 秒内的均值，以得到系数 b. 根据系数对 (a, b)，接收机可以辨别发送机发送的 4 比特.

由于噪声（衰减可以纠正）的存在，接收机收到的值 $\boldsymbol{Y} = (Y_1, Y_2)$ 通常不会与星座

图上的点完全对应，如图所示．这时候，接收机会在星座图中找到距离点 $\boldsymbol{Y}$ 最近的点并读出相应的比特信息．

由于发射机的传输能量有限，$|a|$和$|b|$的值是有界的．因此，如果星座图中有更多数据点（也就是 k 值更大的时候），它们之间的间隔就会更小．这种情况会增加噪声误导接收机的可能性，所以应当根据噪声的能级选择星座图的尺寸．事实上，实际系统正是这么做的．比如，一个有限调制解调器和一个非对称数字用户线路（ADSL）调制解调器会将频带分为小的信道，然后通过度量每个信道的噪声能级为其选择合适的星座图．

多维情况下的MLE

下面对调制、解调，以及为补偿衰减和噪声影响所进行的信号放大作出总结．发射机发送图 5-11 中所示 16 个向量中的 1 个 $\boldsymbol{x}_k = (a_k, b_k)$．我们将发射机传输的向量记为 $\boldsymbol{X}$，将接收机做解调后得到的向量记为 $\boldsymbol{Y}$．

首先假设

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{Z},$$

这里 $\boldsymbol{Z} = (Z_1, Z_2)$且 Z_1、Z_2是独立同分布的服从 $N(0, \sigma^2)$的随机变量，即假设 Y_1、Y_2中的误差为独立的高斯噪声．这种情况下，可以按下述方法计算给定 $\boldsymbol{X}$ 时 $\boldsymbol{Y}$ 的条件分布密度函数 $f_{Y|X}[\boldsymbol{y}|\boldsymbol{x}]$．给定 $\boldsymbol{X} = \boldsymbol{x}$，有 $Y_1 = x_1+Z_1$，$Y_2 = x_2+Z_2$．由于 Z_1 和 Z_2 是独立的，所以 Y_1 和 Y_2 也是独立的．而且，$Y_1 = N(x_1, \sigma^2)$，$Y_2 = N(x_2, \sigma^2)$．因此，

$$f_{Y|X}[\boldsymbol{y} \mid \boldsymbol{x}] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(y_1 - x_1)^2}{2\sigma^2}\} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(y_2 - x_2)^2}{2\sigma^2}\}.$$

前面提到过，$MLE[\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y}]$是使表达式最大的 $\boldsymbol{x} \in \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{16}\}$的值．因此，MLE 得到的值 $\boldsymbol{x}_k$ 使下式最小：

$$\|\boldsymbol{x}_k - \boldsymbol{y}\|^2 = (x_1 - y_1)^2 + (x_2 - y_2)^2.$$

由此可以看到，$MLE[\boldsymbol{X}|\boldsymbol{Y}]$给出的点就是星座图中最接近 $\boldsymbol{Y}$ 值的点．

5.5　假设检验

尽管 MAP 和 MLE 有广泛的应用，但是在设计警报系统、医学检查、错误检测算法和其他一些应用中却难以令人满意．下面介绍一个很重要的规范化问题——假设检验问题．

5.5.1 规范化问题

假设 $X\in\{0, 1\}$，并且已知 Y 在给定 X 下的分布．要解决下面的问题：

$$\text{最大化 } PCD := P[\hat{X}=1\,|X=1]$$

$$\text{服从 } PFA := P[\hat{X}=1\,|X=0]\leqslant\beta.$$

这里，PCD 代表正确检测概率，也就是当 X 为 1 时检测结果为 $X=1$ 的概率．PFA 是错误预警概率，也就是当 X 实际为 0 时检测结果为 $X=1$ 的概率．常数 β 是系统规定的错误预警概率的上界．

我们可以这样直观地理解上述的术语：想象一下 $X=1$ 代表房子着火了．这种情况下，为 $X=1$ 事先假定一个先验概率并不合理，所以无法采用基于 MAP 的规范化问题来寻找答案．同样，MLE 相当于假设 $P(X=1)=1/2$，在这里也不适用．在假设检验的规范化问题下，目标是要以最大的概率检测火灾，但同时保证错误预警概率在预定的范围内，即希望火灾检测器尽可能敏锐，但不至于导致大量错误预警．

由上引出下面的概念．

定义 5.5　ROC

如果问题的解为 $PCD=R(\beta)$，函数 $R(\beta)$就叫作 ROC（Receiver Operating Characteristic，**受试者工作特征**）. ◊

图 5-12 展现了一条典型的 ROC. 这个名称的由来是因为这个函数取决于给定 $X=0$ 和 $X=1$ 时关于 X 的接收信号（即 Y）的条件分布．

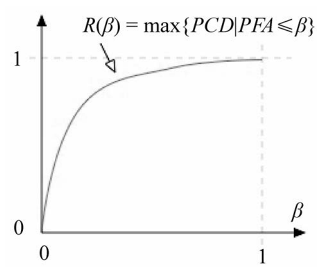


图 5-12　ROC 是当 $PFA\leqslant\beta$ 时正确检测概率的最大值 $R(\beta)$，这里 $R(\beta)$是错误预警概率界限 β 的函数

5.5.2 解答

假设检验问题的解答在下面的定理中阐明．

定理 5.6　奈曼—皮尔逊 [Neyman & Pearson，1933]

假设检验问题的最优解由下式给定

$$\hat{X}=\begin{cases}1, & \text{若}L(Y)>\lambda\\ 1\text{的概率为}\gamma, & \text{若}L(Y)=\lambda\\ 0, & \text{若}L(Y)<\lambda.\end{cases}\tag{5.3}$$

在表达式中，

$$L(y)=\frac{f_{Y|X}[y|1]}{f_{Y|X}[y|0]}$$

为似然比，即当 $X=1$ 时出现 y 的可能性除以当 $X=0$ 时出现 y 的可能性. 而且，$\lambda>0$ 和 $\gamma\in[0,1]$的取值使得 $\hat{X}$ 满足

$$P[\hat{X}=1|X=0]=\beta.$$ □



图 5-13　耶日・奈曼（1894—1981）

可以看到，若 $L(Y)$较大，则 $\hat{X}=1$. 这个结论比较直观，因为 $L(Y)$较大意味着症状值 Y 在 $X=1$ 时比在 $X=0$ 时更可能发生. 在这种情况下，我们更倾向于判定 $X=1$，也就是猜测 $\hat{X}=1$. 当 $L(Y)$较小时，情况是类似的. 通过调整 λ 可以调整检测器的灵敏度. 如果 λ 较小，则检测器更倾向于选择 $\hat{X}=1$，这增加了正确检测的概率，但同时增大了错误预警的概率. 选择 λ 使探测器恰好足够灵敏，使得 $PFA=\beta$. 在某些问题中，也许需要避免让 λ 过于接近临界值. 这一点会在后面的例子中进行解释.

我们会在下一章证明定理 5.6. 现在先来看几个例子.

5.5.3　示例

1. 高斯信道

前面定义了标量的高斯信道模型：

$$Y = X + Z.$$

这里，$Z = N(0, \sigma^2)$并且和 X 无关．在这个模型中，$X \in \{0, 1\}$．接收机希望从收到的信号 Y 中判别出传输信号 X.

下面来看规范化问题 MLE 和 MAP．在 MLE 中，我们希望找到使得 Y 最有可能发生的 X 值．也就是，

$$MLE[X|Y = y] = \arg\max_x f_{Y|X}[y \mid x].$$

当 $Y < 0.5$ 时，结果是 $MLE[X|Y] = 0$，否则 $MLE[X|Y] = 1$.

MAP 会在给定 Y 的情况下，找到 X 在$\{a, b\}$内最可能的值．也就是，

$$MAP[X|Y = y] = \arg\max_x P[X = x \mid Y = y].$$

为了计算 MAP，需要知道 $X = 0$ 的先验概率 p_0．经过计算发现，当 $y \geqslant 0.5 + \sigma^2 \log(p_0 / p_1)$ 时，$MAP[X|Y = y] = 1$，否则 $MAP[X|Y = y] = 0$.

在假设检验的规范化问题中，错误预警概率 $PFA = P[\hat{X} = 1|X = 0]$不能大于 β．根据定理 5.6，要得到最优的判定，需要计算似然比 $L(Y)$：

$$L(y) = \frac{\exp\{-\frac{(y-1)^2}{2\sigma^2}\}}{\exp\{-\frac{y^2}{2\sigma^2}\}} = \exp\{\frac{2y-1}{2\sigma^2}\}.$$

注意，对于任意给定的 λ，都有 $P(L(Y) = \lambda) = 0$．并且，$L(y)$是 y 的严格递增函数．因此，式（5.3）可以简化为

$$\hat{X} = \begin{cases} 1, & 若 y \geqslant y_0 \\ 0, & 其他情况. \end{cases}$$

选取 y_0 使得 $PFA = \beta$，即

$$P[\hat{X} = 1|X = 0] = P[Y \geqslant y_0|X = 0] = \beta.$$

如果 $X = 0$，则 $Y = N(0, \sigma^2)$．因此，y_0 的值应当满足

$$P(N(0, \sigma^2) \geqslant y_0) = \beta.$$

即使得

$$P(N(0, 1) \geqslant \frac{y_0}{\sigma}) = \beta.$$

比如，图 3-7 展示了当 $\beta = 5\%$时，$y_0/\sigma = 1.65$．图 5-14 展示了解的性质．

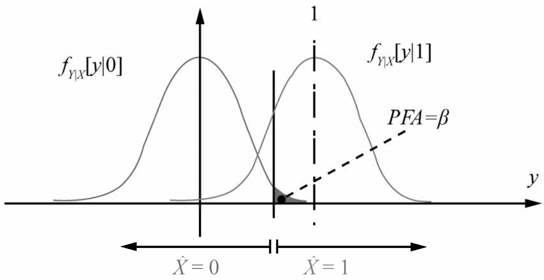


图 5-14　某一高斯信道假设检验问题的解

现在计算高斯信道的 ROC．令 $y(\beta)$使得 $P(N(0, 1)\geqslant y(\beta)) = \beta$．于是 $y_0 = y(\beta)\sigma$．在这种情况下，正确检测的概率为

$$\begin{aligned}PCD &= P[\hat{X}=1 \mid X=1] = P[Y \geqslant y_0 \mid X=1] = P(N(1,\sigma^2) \geqslant y_0)\\ &= P(N(0,\sigma^2) \geqslant y_0 - 1) = P(N(0,1) \geqslant \sigma^{-1}y_0 - \sigma^{-1})\\ &= P(N(0,1) \geqslant y(\beta) - \sigma^{-1}).\end{aligned}$$

图 5-15 中显示了不同 σ 的 ROC（由 Matlab 得到）．可以看到，系统的性能随着噪声的增强而减弱．

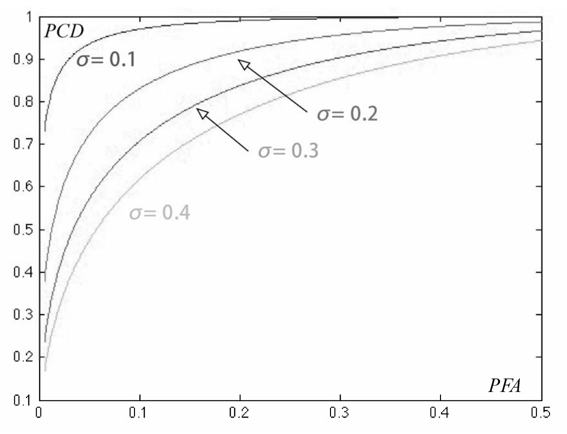


图 5-15　高斯信道 $Y = X+Z$ 的 ROC，其中 $X\in \{0, 1\}$，$Z = N(0, \sigma^2)$

2. 指数分布随机变量的均值

在第二个例子中，测试指数分布随机变量的均值．问题的背景是一台生产灯泡的机器，它生产的灯泡寿命服从指数分布．机器有两个状态 0 和 1．当状态 $X = x\in \{0, 1\}$ 时，平均寿命为 $1/\lambda_x$（假设 $\lambda_0 < \lambda_1$）．这台机器在 $X = 1$ 时会使产品受到影响，即灯泡的寿命缩短．

令 $Y=(Y_1,\ \cdots,\ Y_n)$表示 n 个灯泡的寿命．我们希望检测到 $X=1$，并使错误预警概率 $PFA\leqslant\beta=5\%$.

首先计算

$$L(y)=\frac{f_{Y|X}[y\,|\,1]}{f_{Y|X}[y\,|\,0]}=\frac{\prod_{i=1}^{n}\lambda_1\exp\{-\lambda_1 y_i\}}{\prod_{i=1}^{n}\lambda_0\exp\{-\lambda_0 y_i\}}$$
$$=(\frac{\lambda_1}{\lambda_0})^n\exp\{-(\lambda_1-\lambda_0)\sum_{i=1}^{n}y_i\}.$$

因为 $\lambda_0<\lambda_1$，$L(y)$是 $\sum_i y_i$ 的严格递减函数，并且对任何 λ，都有 $P(L(Y)=\lambda)=0$．所以，式（5.3）可简化为

$$\hat{X}=\begin{cases}1, & 若\sum_{i=1}^{n}Y_i\leqslant a\\ 0, & 其他情况.\end{cases}$$

这里 a 的值使得

$$P[\sum_{i=1}^{n}Y_i\leqslant a\,|\,X=0]=\beta=5\%.$$

当 $X=0$ 时，Y_i是均值为 $1/\lambda_0$ 的独立同分布指数随机变量．它们和的分布很复杂（伽马分布），所以用中心极限定理来估计.

我们有①

$$\frac{Y_1+\cdots+Y_n-n\lambda_0^{-1}}{\sqrt{n}}\approx N(0,\lambda_0^{-2}).$$

因此，

$$\sum_{i=1}^{n}Y_i\leqslant a\Leftrightarrow\frac{Y_1+\cdots+Y_n-n\lambda_0^{-1}}{\sqrt{n}}\leqslant\frac{a-n\lambda_0^{-1}}{\sqrt{n}}.$$

所以，

$$P[\sum_{i=1}^{n}Y_i\leqslant a\,|\,X=0]\approx P(N(0,\lambda_0^{-2})\leqslant\frac{a-n\lambda_0^{-1}}{\sqrt{n}})$$
$$=P(N(0,1)\leqslant\lambda_0\frac{a-n\lambda_0^{-1}}{\sqrt{n}}).$$

因此，如果要让这个概率等于 5%，由式（3.2），一定要选取 a 的值使得

$$\lambda_0\frac{a-n\lambda_0^{-1}}{\sqrt{n}}=1.65,$$

① 注意，$\mathrm{var}(Y_i)=\lambda_0^{-2}$.

即

$$a = (n+1.65\sqrt{n})\lambda_0^{-1}.$$

在这个例子里需要注意的一点是，$\hat{X}$ 的计算是基于 $Y_1+\cdots+Y_n$ 的．所以，虽然测量了 n 个灯泡的寿命，但是最终起决定作用的只有它们的和，即它们的均值．

3．硬币的偏差

在第三个例子中，观察的对象是 n 个硬币的抛掷结果．假设给定 $X = x\in\{0, 1\}$，硬币是独立同分布的 $B(p_x)$．也就是说，对给定的 $X = x$，抛硬币的结果 $Y_1, \cdots, Y_n$ 是独立同分布的，等于 1 的概率为 p_x，否则等于 0．假设 $p_1 > p_0 = 0.5$，这个问题等同于检测硬币是均匀的还是有偏差的．

这里，随机变量 Y_i 是离散的．我们有

$$P[Y_i = y_i, i = 1,\cdots,n \mid X = x] = \prod_{i=1}^{n} p_x^{Y_i}(1-p_x)^{1-Y_i}$$
$$= p_x^{S}(1-p_x)^{n-S}，\text{其中} S = Y_1+\cdots+Y_n.$$

所以，

$$L(Y_1,\cdots,Y_n) = \frac{P[Y_i = y_i, i = 1,\cdots,n \mid X = 1]}{P[Y_i = y_i, i = 1,\cdots,n \mid X = 0]}$$
$$= \left(\frac{p_1}{p_0}\right)^{S}\left(\frac{1-p_1}{1-p_0}\right)^{n-S} = \left(\frac{1-p_1}{1-p_0}\right)^{n}\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^{S}.$$

由于 $p_1 > p_0$，似然比随 S 增加而增加．所以，假设检验问题的解为

$$\hat{X} = 1\{S \geqslant n_0\},$$

这里 n_0 的值使得 $P[S \geqslant n_0 \mid X = 0] \approx \beta$. 为了计算 n_0, 在 $X = 0$ 时用中心极限定理估计 S:

$$P[S \geqslant n_0 \mid X = 0] = P\left(\frac{S-np_0}{\sqrt{n}} \geqslant \frac{n_0-np_0}{\sqrt{n}} \mid X = 0\right)$$
$$\approx P\left(N(0, p_0(1-p_0)) \geqslant \frac{n_0-np_0}{\sqrt{n}}\right)$$
$$= P\left(N(0, 0.25) \geqslant \frac{n_0-np_0}{\sqrt{n}}\right) = P\left(N(0,1) \geqslant \frac{2n_0-n}{\sqrt{n}}\right).$$

如果 $\beta = 5\%$，那么由式（3.2）可得

$$\frac{2n_0-n}{\sqrt{n}} = 1.65.$$

因此，

$$n_0 = 0.5n+0.83\sqrt{n}.$$

4．离散观测

在以上的例子中，$L(Y)$都是连续的．在这样的情况下，$L(Y) = \lambda$ 的概率总是 0，

所以没有必要在 Y 取某些特定值的时候随机地选择 $\hat{X}$．现在来看一个需要随机选择的例子．

仍然考虑判定 $\hat{X}$ 为 0 还是 1 的问题. 我们希望以最大化正确检测的概率并保证错误预警概率不大于 β．如果不作任何观测，那么最优的做法是以概率 β 判定 $\hat{X}=1$．这样错误预警的概率得到了保证，同时正确检测的概率也为 β．这结果优于直接判定 $\hat{X}$ 总为 0．

下面考虑一个更复杂的例子． $Y \in \{A, B, C\}$，并且

$$P[Y=A|X=1]=0.2，P[Y=B|X=1]=0.2，P[Y=C|X=1]=0.6$$

$$P[Y=A|X=0]=0.2，P[Y=B|X=0]=0.5，P[Y=C|X=0]=0.3.$$

那么似然比 $L(y)=P[Y=y|X=1]/P[Y=y|X=0]$的值为

$$L(A)=1，L(B)=0.4，L(C)=2.$$

现在，依照 L 的值把观测到的结果按升序排列．结果如图 5-16 所示．

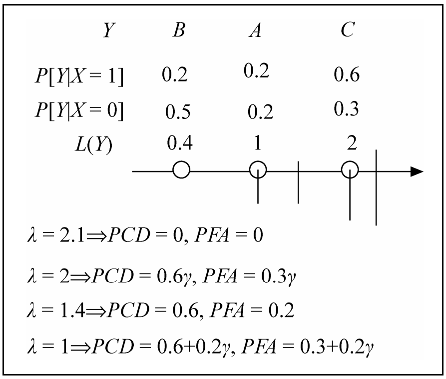


图 5-16　3 个可能的观测结果

因此，这一假设检验的结果可以通过寻找阈值 λ 与随机概率 γ 来满足下式：

$$P[\hat{X}=1 \mid Y]=1\{L(Y)>\lambda\}+\gamma 1\{L(Y)=\lambda\}.$$

同时，λ 与 γ 的取值也使得错误预警概率为 β.

图 5-16 显示，如果选择 $\lambda=2.1$，对于所有 Y 的取值，$L(Y)<\gamma$. 因此总是判定 $\hat{X}=0$. 相应地，正确的检测概率和错误预警概率也都为 0．图 5-16 还告诉我们，如果选择 $\lambda=2$ 和参数 γ，那么在 $L(Y)=2$ 的时候我们会以概率 γ 判定 $\hat{X}=1$．由于如果 $X=0$，则 $Y=C$ 的概率为 0.3 并且 $L(Y)=2$（意味着我们以概率 γ 选取 $\hat{X}=1$），该方法判定 $\hat{X}=1$ 的概率为 0.3γ．而这恰恰是定理 5.6 的结论．

我们还能从图里读出许多其他结论．如果将 λ 从 2.1 减小到 0.39，那么犯错的概率会从 0 增加到 1．另外，我们总能通过调整 λ 的值以及选择相应的 γ 来满足任意[0, 1]

内的错误预警概率. 比如, 可以选择 $\lambda=2$ 且 $\gamma=0.05/0.3$ 以保证 $PFA=0.05$. 类似地, 选择 $\lambda=1$ 并且 $\gamma=0.5$ 能保证 $PFA=0.4$. 在此情况下, 如果观测值 $Y=C$, 则判定 $\hat{X}=1$; 当 $Y=A$ 时, 以概率 0.5 判定 $\hat{X}=1$. 这么一来, 当 $X=0$ 的时候, $PFA=0.3+0.2\times0.5=0.4$. 同时可以看到 $PCD=0.6+0.2\times0.5=0.7$.

图 5-17 给出了 PCD 与 PFA 的关系.

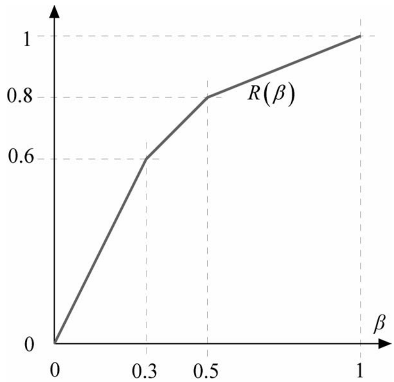


图 5-17 离散例子中的受试者工作特征

5.6 小结

- ❑ MAP 和 MLE
- ❑ BPSK
- ❑ 霍夫曼编码
- ❑ 独立高斯误差
- ❑ 假设检验：奈曼–皮尔逊定理

重要方程与公式

贝叶斯准则	$\pi_i = p_i q_i / \Sigma_j p_j q_j$	定理5.1
$MAP[X\|Y=y]$	$\arg\max_x P[X=x\|Y=y]$	定义5.2
$MLE[X\|Y=y]$	$\arg\max_x P[Y=y\|X=x]$	定义5.2
似然比	$L(y)=f_{Y\|X}[y\|1]/f_{Y\|X}[y\|0]$	定理5.6
高斯信道	$MAP[X\mid Y=y]=1\{y\geqslant\frac{1}{2}+\sigma^2\log(\frac{p_0}{p_1})\}$	式（5.2）
奈曼–皮尔逊定理	$P[\hat{X}=1\|Y]=1\{L(Y)>\lambda\}+\gamma 1\{L(Y)=\lambda\}$	定理5.6
ROC	$ROC(\beta)=\max.PCD$ 使得 $PFA\leqslant\beta$	定义5.5

5.7 参考资料

检测理论是一个很经典的理论，它是数字通讯的核心（可以参见 Proakis 的 *Digital Communications, 4th Edition*）．奈曼–皮尔逊定理在“On the Problem of the most Efficient Tests of Statistical Hypotheses”, 1933 中有介绍．对于假设检验的讨论，可见 Lehmann 的 *Testing Statistical Hypotheses*, *3rd Edition*．如果读者对于数字通信（特别是无线通信）的细节感兴趣，请参看 Tse 和 Viswanath 的 *Fundamentals of Wireless Communication*．

5.8 练习

1. 假设当 $X=0$ 时，$Y=\mathcal{N}(0,1)$；当 $X=1$ 时，$Y=\mathcal{N}(0,\sigma^2)$，其中 $\sigma^2>1$．计算 $MLE[X|Y]$．

2. 令 X 和 Y 为独立且满足分布 $U[0,1]$的随机变量．定义 $V=X+Y$ 和 $W=X-Y$．

(a) 证明 V 和 W 不相关．

(b) V 和 W 是否独立？证明之．

3. 一个数据链路采用了如图 5-11 所示的 16QAM 星座图，并有 $\boldsymbol{x}_1=(1,\ -1)$．设接收机接收到的信号为 $\boldsymbol{Y}=\boldsymbol{X}+\boldsymbol{Z}$，这里 $\boldsymbol{Z}=_{\boldsymbol{D}}\mathcal{N}(\boldsymbol{0},\sigma^2\boldsymbol{I})$．接收机运用最大后验机制．用 Matlab 仿真这一系统来估计 $\sigma=0.2$ 和 0.3 时的错误比例． C.5

4. 通过 Matlab 产生独立同分布的 $U[0,1]$随机变量 X_n 来验证中心极限定理，即生成随机变量$\{X_1,\ \cdots,X_N\}$，$N=10\,000$．计算 C.2

$$Y_n=\frac{X_{100n+1}+\cdots+X_{(n+1)100}-50}{10},\ n=0,\ 1,\ \cdots,\ 99.$$

绘出$\{Y_0,\ \cdots,Y_{99}\}$的经验累积分布函数并和 $\mathcal{N}(0,1/12)$的累积分布函数作比较．

5. 检验一个相当于二元对称信道、错误概率 $\varepsilon\in[0,0.5)$的数字链路．

(a) 观测链路的输入输出，如何得出 ε 的最大似然估计？

(b) 已知输入是独立同分布的比特，等于 1 的概率为 0.6，等于 0 的概率为 0.4．观测 n 个输出后，如何计算出 ε 的最大似然估计？

(c) 和上面情况相同，不过已知 ε 在$[0,\ 0.5)$上的概率密度函数为 $4-8x$．对于给定的 n 个输出，如何计算出 ε 的最大后验概率？

6. 和上一题情况相同，观测二元对称信道的输入输出．通过假设检验的方法判定 $\varepsilon>0.1$ 的概率，并保证错误预警概率不超过 5%．假定 n 非常大，并且运用中心极限定理．

7. 随机变量 X 满足 $P(X=1)=2/3$，$P(X=0)=1/3$．当 $X=1$ 时，随机变量 Y 服从参数为 1 的指数分布．当 $X=0$ 时，随机变量 Y 在$[0,\ 2]$上均匀分布．（提示：注意 $Y>2$

的情况.）

(a) 计算 $MLE[X|Y]$.

(b) 计算 $MAP[X|Y]$.

(c) 解决下述假设检验问题：最大化 $P[\hat{X}=1|X=1]$且满足 $P[\hat{X}=1|X=0]\leqslant 5\%$.

8. 对下面的通信信道进行仿真. 有一个独立同分布的源，根据先验分布 $\pi=[p_1, p_2, p_3, p_4]$生成符号$\{1, 2, 3, 4\}$. 这些符号根据四相相移键控机制调制，即被映射到星座图$(\pm1, \pm1)$上. 通信建立在基带高斯信道上：如果发送的信号是(x_1, x_2)，则接收的信号是

C.5

$$y_1 = x_1+Z_1,\quad y_2 = x_2+Z_2,$$

这里的 Z_1 和 Z_2 是独立的满足 $N(0, \sigma^2)$分布的随机变量. 找出最大后验估计和最大似然估计.

运用 Simulink（推荐）或者 Matlab 模拟 $\pi=[0.1, 0.2, 0.3, 0.4]$，$\sigma=0.1$ 和 $\sigma=0.5$ 情况下的信道. 评估正确检测的概率.

9. 假设 X 等概率地取$\{1, 2, 3\}$中的值. 对于给定的 X，随机变量 Y 满足分布 $\mathcal{N}(X, 1)$.

(a) 计算 $MAP[X|Y]$.

(b) 计算 $MLE[X|Y]$.

(c) 计算 $E((X-Y)^2)$.

10. 随机变量 X 满足 $P(X=0)=P(X=1)=0.5$. 对给定的 X，随机变量 Y_n 是满足分布 $U[0, 1.1-0.1X]$的独立同分布随机变量. 我们希望从观测值 Y_n 中判别出 $\hat{X}$. 每个观测的代价是 $\beta>0$. 为了得到好的数值解，假设

$$\beta = 0.018 \approx 0.5(1.1)^{-10}\log(1.1).$$

(a) 假设观测到 $Y^n=(Y_1, \cdots Y_n)$. 基于此，如何最大化 $\hat{X}=X$ 的概率？

(b) 此时对应的 $P(\hat{X}_n=X)$为多少？

(c) 选择使 $P(\hat{X}_n=X)-\beta_n$ 最大的 n，其中 $\hat{X}_n$ 选自基底$(Y_1, \cdots, Y_n)$. 提示：使用

$$\frac{\mathrm{d}}{\mathrm{d}x}(a^x) = a^x\log(a).$$

11. 随机变量 X 服从参数为 1 的指数分布. 对于给定的 X，随机变量 Y 为参数为 X 的指数变量.

(a) 计算 $MLE[X|Y]$.

(b) 计算 $MAP[X|Y]$.

(c) 解决下述假设检验问题：最大化 $P[\hat{X}=1|X=a]$且满足 $P[\hat{X}=1|X=1]\leqslant 5\%$，其中 $a>1$ 为给定常数.

12. 考虑参数为 θ 的指数分布随机变量 Y. 观察到这个随机变量的 n 个独立同分布随机样本 $Y_1, \cdots, Y_n$. 计算 $\hat{\theta}=MLE[\theta|Y_1, \cdots, Y_n]$. 这个估计的偏差 $E[\hat{\theta}-\theta|\theta]$是多少？

当 n 趋向于无穷时，这一偏差是否趋于 0?

13. 假设 $Y = {}_DU[a, b]$. 观察到这个随机变量的 n 个独立同分布随机样本 $Y_1, \cdots, Y_n$. 计算 a 和 b 的最大似然估计 $\hat{a}$ 和 $\hat{b}$. $\hat{a}$ 和 $\hat{b}$ 的偏差是多少?

14. 考虑一个假设检验问题. 设 X 和 $\hat{X}$ 在 $\{0, 1\}$ 中取值. $\hat{X}$ 的取值取决于随机向量 $\boldsymbol{Y}$ 的观测值. 假设对于给定 $X = i$, $\boldsymbol{Y}$ 的概率密度为 $f_i(\boldsymbol{y})$, 其中 $i = 0, 1$, 而且定义 $L(\boldsymbol{y}) := f_1(\boldsymbol{y})/f_0(\boldsymbol{y})$.

定义 $g(\beta)$为满足 $P[\hat{X} = 1|X = 0] \leqslant \beta$ 时, $P[\hat{X} = 1|X = 1]$的最大值, 其中 $\beta \in [0, 1]$. 那么（如果有，选择正确答案）:

- ❑ $g(\beta) \geqslant 1-\beta$;
- ❑ $g(\beta) \geqslant \beta$;
- ❑ 最佳决策是由函数 $h(\boldsymbol{y}) = P[\hat{X} = 1|\boldsymbol{Y} = \boldsymbol{y}]$描述的，并且该函数对于 $f_1(\boldsymbol{y})/f_0(\boldsymbol{y})$是非减的.

15. 给定 $\theta \in [0, 1]$, $\boldsymbol{X} = \theta(1, 1)' + \boldsymbol{V}$, 这里 V_1 和 V_2 是独立的, 在$[-2, 2]$上均匀分布. 解决假设检验问题：最大化 $P[\hat{\theta} = 1|\theta = 1]$使得 $P[\hat{\theta} = 1|\theta = 0] \leqslant 5\%$.

16. 给定 $\theta = 1$, $X = {}_DExp(1)$和给定 $\theta = 0$, $X = {}_DU[0, 2]$.

(a) 找出 $\hat{\theta} = HT[\theta|X, \beta]$, 即随机变量 $\hat{\theta}$ 是 X 的一个函数，使得在满足 $P[\hat{\theta} = 1|\theta = 0] \leqslant \beta$ 时 $P[\hat{\theta} = 1|\theta = 1]$最大.

(b) 计算 $\alpha(\beta) = P[\hat{\theta} = 1|\theta = 1]$的结果.

(c) 描出受试者工作特征曲线 $\alpha(\beta)$, 其中 $\beta \in [0, 1]$.

17. 观测一个随机序列$\{X_n, n = 0, 1, 2, \cdots\}$. $\theta = 0$ 且该序列是独立同分布的伯努利序列, $P(X_n = 0) = P(X_n = 1) = 0.5$ 的概率为 P; $\theta = 1$ 且该序列是在$\{0, 1\}$上的稳定马尔可夫链，转移概率 $P(0, 1) = P(1, 0) = \alpha$ 的概率为 $1-P$. 参数 α 为$(0, 1)$上的给定值.

(a) 计算 $MAP[\theta|X_0, \cdots, X_n]$.

(b) 讨论 $\hat{\theta}_n$ 的收敛性.

(c) 讨论组合假设检验问题：当 $\theta = 1$ 时, $\alpha < 0.5$; 当 $\theta = 0$ 时, $\alpha = 0.5$.

18. 如果 $\theta = 0$, 序列$\{X_n, n \geqslant 0\}$是定义在有限集 $\mathcal{X}$ 上的马尔可夫链，转移矩阵为 P_0. 如果 $\theta = 1$, 转移矩阵为 P_1. 在这两种情况下都有 $X_0 = x_0$. 计算 $MLE[\theta|X_0, \cdots, X_n]$.

第6章

数字链路—B

主题：霍夫曼编码的最优性，LDPC 码，奈曼–皮尔逊定理的证明，联合高斯随机变量

背景知识：第 5 章

第 5 章介绍了检测和假设检验问题，霍夫曼编码，以及独立高斯错误的情况．在本章，我们会证明霍夫曼编码的最优性，证明奈曼–皮尔逊定理，并解释 LDPC 码和联合高斯随机变量．

6.1 霍夫曼编码最优性的证明

第 5 章给出了下面的定理，本节会给出证明．

定理 6.1　霍夫曼编码的最优性

对于每个字符，霍夫曼编码所需的平均比特数在所有无前缀编码中最小．

□

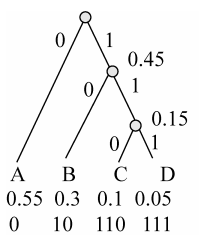


图 6-1　霍夫曼编码

在霍夫曼 1952 年发表的文章中，他通过对字符数采用数学归纳法论证了该定理．首先假设霍夫曼编码在字符数为 n 的情况下所采用的平均比特数最少．记该比特数为 $L(n)$．现在假设存在另一编码树 T，在字符数为 $n+1$ 的时候使用的平均比特数为

$A(n+1)$，比霍夫曼编码更短．令 X 和 Y 为两个使用频率最少的字符，记它们的频率为 $x\geqslant y$．可以在 T 中选这样的字符，使其路径长度最大并且 Y 在 T 中的路径最长．也可以将 Y 与 T 中一个更常出现的字符进行交换，就能减少平均路径长度．

接下来，先假定也可以选择 X 和 Y，使其在 T 中为兄弟节点（下面会证明）．那么通过将 X 和 Y 合并成一个父节点 Z（频率为 $z=x+y$），就建立了对 n 个字符的编码，其平均路径长度为 $A(n+1)-z$．由于霍夫曼编码对 n 字符的情况为最优，有 $L(n)\leqslant A(n+1)-z$．现在，对于 $n+1$ 个字符的霍夫曼编码，也合并 X 和 Y．这样一来，霍夫曼编码在 $n+1$ 个字符时的平均路径长度为 $L(n+1):=L(n)+z\leqslant A(n+1)$．这与霍夫曼编码在 $n+1$ 个字符时不是最优的假设相矛盾．

接下来只需证明 X 和 Y 可以成为兄弟节点即可．首先，由于 Y 有最长的路径长度，它不可能是唯一的子节点．否则，我们会用 Y 代替它的父节点从而减小其路径长度．假设 Y 有一个不是 X 的兄弟节点 V，交换 V 和 X 不会增加平均路径长度，因为 V 的频率不会比 X 小．这样即完成了论证． ♣

6.2 低密度奇偶校验码（LDPC码）

LDPC（Low Density Parity Check，低密度奇偶校验）码是实际应用中最有效的编码之一．加拉格在 1960 年的论文中发表了这一编码[①]．这些编码在今天的应用十分广泛，比如卫星视频传送．它们对于二元对称信道和其他许多信道几乎能达到最优的性能．



图 6-2 罗伯特 G. 加拉格（1931— ）

LDPC 码如下所示．设 $\boldsymbol{x}\in\{0,1\}^n$ 是一个即将传送的 n 比特字符串，我们将这一字符串用一个 m 比特的 $\boldsymbol{y}$ 进行扩增，这里

$$\boldsymbol{y}=H\boldsymbol{x} \tag{6.1}$$

在式（6.1）里，H 是一个 $m\times n$ 的$\{0,1\}$矩阵，$\boldsymbol{x}$ 和 $\boldsymbol{y}$ 均为列向量，并且加法操作是模 2 加法．比如，如果

且 $\boldsymbol{x}=[01001010]$，那么 $\boldsymbol{y}=[1110]$．这个由 $\boldsymbol{x}$ 计算奇偶校验位 $\boldsymbol{y}$ 的过程可以通过泰纳图（Tanner Graph）表示，如图 6-3 所示．

① R.G. Gallager，"低密度奇偶校验码"，MIT 出版社，1963．

$$H = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

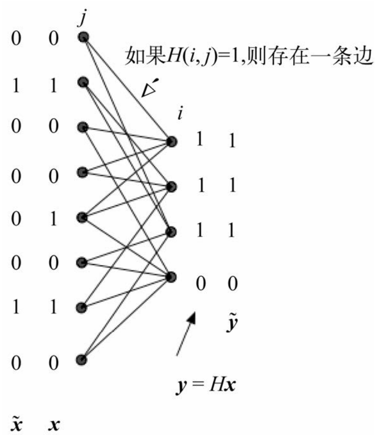


图 6-3　LDPC 码的泰纳图表示．该图指出了 H 阵的非零元素，使得 $\boldsymbol{y} = H\mathbf{x}$．接收机收到 $\tilde{\boldsymbol{x}}$ 和 $\tilde{\boldsymbol{y}}$ 不是 $\boldsymbol{x}$ 和 $\boldsymbol{y}$，点 x_j 叫作消息节点，点 y_i 叫作检验节点

因此，传送的是 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 而不仅仅是比特串 $\boldsymbol{x}$．$\boldsymbol{y}$ 中的比特是奇偶校验位．由于传输过程中可能会有错误，接收机接收到 $\tilde{\boldsymbol{x}}$ 和 $\tilde{\boldsymbol{y}}$ 而不是 $\boldsymbol{x}$ 和 $\boldsymbol{y}$．然后，接收机计算 $H\tilde{\boldsymbol{x}}$ 并将结果与 $\tilde{\boldsymbol{y}}$ 相比较．这一校验方式是，如果 $\tilde{\boldsymbol{y}} = H\tilde{\boldsymbol{x}}$，那么很有可能 $\tilde{\boldsymbol{x}} = \boldsymbol{x}$，$\tilde{\boldsymbol{y}} = \boldsymbol{y}$．它的依据是，很难在发生传输错误之后还能满足 $\tilde{\boldsymbol{y}} = H\tilde{\boldsymbol{x}}$．因此，如果选择恰当的 H 矩阵，那么这一机制应当能很好地检查出错误．

除了检测错误之外，LDPC 码还可用于纠错．如果 $\tilde{\boldsymbol{y}} \neq H\tilde{\boldsymbol{x}}$，那么我们希望通过改变最少数目的比特，使得 $\tilde{\boldsymbol{y}} = H\tilde{\boldsymbol{x}}$ 重新得到满足．这里的依据是，如果比特发生错误的概率很小并且均为独立同分布，那么按照上述方法找出的比特错误就是最有可能发生的传送错误．不过，找到正确元素组合的计算量随着元素个数指数增加．所以在实际中，人们用迭代算法来逼近最终解．

下面介绍一个经常用到的解码算法，叫作置信传播算法（BP）．假设每一个接收到比特的错误率为 $\varepsilon \ll 1$，正确率为 $\bar{\varepsilon} = 1-\varepsilon$，且比特之间相互独立．同时假设发射机传输的比特 x_j 以等概率取 0 或 1．由对称性可知，这意味着校验位 y_i 也以相同概率取 0 或 1．在该算法中，消息节点 x_j 和检验节点 y_i 沿着图 6-3 的链路传递它们对 x_j 等于 1 的概率估值．

在算法的奇数步中，每个节点 x_j 向每一个与之相连接的节点 y_i 发送一个 $P(x_j = 1)$的

估计．节点 y_i 结合收到的估计，向每个 x_j 节点返回一个 $P(x_j = 1)$的新的估值．y 节点进行的运算如下：考虑图 6-4 所示的情况，其中节点 y_1 得到估计值 $a = P(x_1 = 1)$, $b = P(x_2 = 1)$, $c = P(x_3 = 1)$. 假设 $\tilde{y}_1 = 1$，于是节点 y_1 采用贝叶斯准则计算 $P[y_1 = 1| \tilde{y}_1] = 1-\varepsilon = \bar{\varepsilon}$．从图中可以看到 $x_1+x_2+x_3 = y_1$，因此节点 y_1 计算在 $\{x_2, x_3, y_1\}$ 中有奇数个比特等于 1 的概率，并用其作为 $x_1 = 1$ 的概率估计.

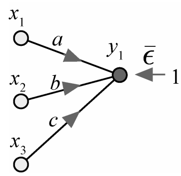


图 6-4　节点 y_1 从 x 节点处获得估计值，再计算新的估计值

下面来看计算的具体步骤．假设 x_1, …, x_n 是独立的{0, 1}随机变量，并记 $p_i = P(x_i = 1)$．注意到，如果在 $\{x_1, \cdots, x_n\}$ 中等于 1 的变量数为偶数，

$$1-(1-2x_1)\times\cdots\times(1-2x_n)$$

等于 0；否则等于 2．对其取期望值可以得到

$$2P(\text{奇数}) = 1-\prod_{i=1}^{n}(1-2p_i),$$

所以，

$$P(\text{奇数}) = \frac{1}{2}-\frac{1}{2}\prod_{i=1}^{n}(1-2p_i). \tag{6.2}$$

因此，在图 6-4 中，

$$\begin{aligned} P(x_1=1) &= P(x_2, x_3, y_1\text{中有奇数个}1) \\ &= \frac{1}{2}-\frac{1}{2}(1-2b)(1-2c)(1-2\bar{\varepsilon}). \end{aligned} \tag{6.3}$$

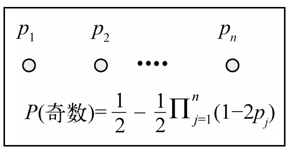


图 6-5　每个节点 j 等于 1 的概率为 p_j，其他情况下等于 0，各个节点相互独立．图中的表达式给出了存在奇数个节点为 1 的概率

图 6-3 中的 y 节点就运用了上述方法来计算新的估计值，并将其发回 x 节点.

在算法的偶数步中，每个 x_j 节点会从 $\tilde{x}_j$ 处以及 y 节点处得到它们在前一步中的估值．这里，节点 x_j 假设它得到的不同估计值来源于独立观察．也就是说，节点 x_j 从独立的专家处，即 $\tilde{x}_j$ 和它在图中相连的 y_i，得到关于 $P(x_j = 1)$的意见．然后，节点 x_j 会

融合这些专家的意见并重新进行估计.

应该如何结合独立的专家观点呢？假设 N 个专家的观察值为 Y_1，…，Y_N，并给出估计 $p_i = P[X = 1|Y_i]$．假定先验概率是 $P(X = 1) = P(X = 0) = 1/2$，采用下面的方法来估计 $P[X = 1|p_1,\cdots,p_N]$：

$$\begin{aligned}P[X=1\mid Y_1,\cdots,Y_N] &= \frac{P(X=1,Y_1,\cdots,Y_N)}{P(Y_1,\cdots,Y_N)} \\ &= \frac{P[Y_1,\cdots,Y_N\mid X=1]P(\mathrm{X}=1)}{\sum_{x=0,1}P[Y_1,\cdots,Y_N\mid X=x]P(X=x)} \\ &= \frac{P[Y_1\mid X=1]\times\cdots\times P[Y_N\mid X=1]}{\sum_{x=0,1}P[Y_1\mid X=x]\times\cdots\times P[Y_N\mid X=x]}.\end{aligned} \quad (6.4)$$

现在计算 $P[Y_n|X = x]$，

$$P[Y_n\mid X=x]=\frac{P(X=x,Y_n)}{P(X=x)}=\frac{P[X=x\mid Y_n]P(Y_n)}{1/2}.$$

因此，

$$P[Y_n\mid X=0]=2(1-p_n)P(Y_n)\text{ 且 }P[Y_n\mid X=1]=2p_nP(Y_n).$$

将这些表达式带入式（6.4）可得

$$P[X=1\mid Y_1,\cdots,Y_N]=\frac{p_1\cdots p_N}{p_1p_2\cdots p_N+(1-p_1)\cdots(1-p_N)}. \quad (6.5)$$

图 6-6 展示了这一过程.

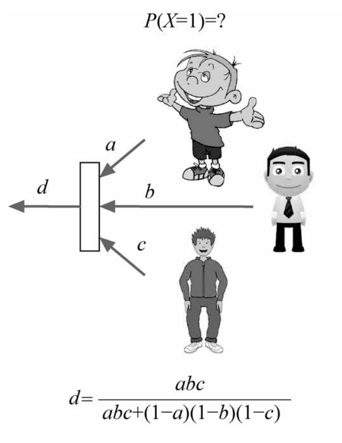


图 6-6　当先验概率是 1/2 时，结合独立专家对 $P(X = 1)$的观点

将这一结果运用到图 6-7 所示的情况．在图中，节点 x_1 通过观察 $\tilde{x}_1 = 0$ 得到对 $P(x_1 = 1)$的估计 ε．它同时从节点 y_1, y_2, y_3 处得到估计值 a, b, c，并假设这些估计相互独立.

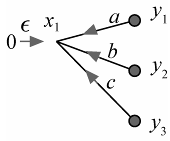


图 6-7　节点 x_1 从节点 y 处得到 $P(x_1 = 1)$的估计，并计算新的估计值

为了计算新的估计值，节点 x_1 结合从 $\tilde{x}_1, y_2, y_3$ 处得到的估计值

$$\frac{\varepsilon bc}{\varepsilon bc + \bar{\varepsilon}\bar{b}\bar{c}}. \tag{6.6}$$

其中 $\bar{b} = 1-b$, $\bar{c} = 1-c$. 在下一步中，节点 x_1 把这一估计值发给节点 y_1. 它也会计算出节点 y_2 和 y_3 的估计值.

总结起来，该算法的步骤如下. 在奇数步时，节点 x_j 发送 $X(i, j)$给每个节点 y_i. 在每个偶数步，节点 y_i 发送 $Y(i, j)$给每个节点 x_j:

$$Y(i,j) = \frac{1}{2} - \frac{1}{2}(1-2\varepsilon)(1-2\bar{y}_i)\prod\nolimits_{s\in A(i,j)}(1-2X(i,s)) , \tag{6.7}$$

其中 $A(i,j) = \{s \neq j \mid H(i,s) = 1\}$ 且

$$X(i,j) = \frac{N(i,j)}{N(i,j)+D(i,j)}. \tag{6.8}$$

这里,

$$N(i,j) = P[x_j = 1 \mid \tilde{x}_j]\prod\nolimits_{\{v\neq i \mid H(v,j)=1\}} Y(v,j),$$

并且

$$D(i,j) = P[x_j = 0 \mid \tilde{x}_j]\prod\nolimits_{\{v\neq i \mid H(v,j)=1\}} (1-Y(v,j)),$$

$$P[x_j = 1 \mid \tilde{x}_j] = \varepsilon + (1-2\varepsilon)\tilde{x}_j .$$

同时，节点 x_j 通过下面的方法结合专家观点并更新 x_j 等于 1 的概率:

C.6

$$X(j) = \frac{N(j)}{N(j)+D(j)} , \tag{6.9}$$

其中,

$$N(j) = P[x_j = 1 \mid \tilde{x}_j]\prod\nolimits_{\{v \mid H(v,j)=1\}} Y(v,j) ,$$

并且

$$D(j) = P[x_j = 0 \mid \tilde{x}_j]\prod\nolimits_{\{v \mid H(v,j)=1\}} (1-Y(v,j)).$$

在经过足够多次的迭代后，可以得到检验决策 $x_j = 1\{X(j) \geqslant 0.5\}$.

图 6-8 展示了随时间演变的 x_j 等于 1 的估计概率. 我们的代码直接采用了本节的

公式．置信传播算法还有更复杂的实现方式，使用 log 函数的和而非概率的乘积．

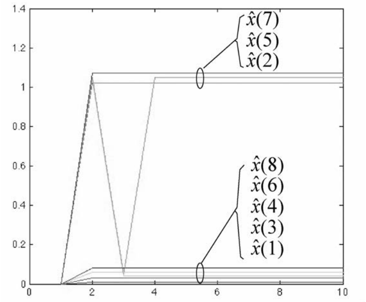


图 6-8　置信传播算法应用于图 6-3 展示的系统．横轴是算法的步数，竖轴是在该步得到的 $x(i)$最佳猜测．为清楚起见，我们将预测的最小间隔设为 0.1．最终的检测是[0, 1, 0, 0, 1, 0, 1, 0]．可以直观地看到这是最佳猜测（另见彩插）

大量的仿真和深层的理论表明，这个算法在图没有小环时表现很好．在这种情况下，算法中的估计值几乎可以被认为是从独立的观测中得到的．

6.3　联合高斯分布随机变量

在很多系统中，向量 $\boldsymbol{Y}$ 中每个分量上的错误并不是独立的．一个适用于这种情况的模型如下：

$$\boldsymbol{Y} = \boldsymbol{X}+A\boldsymbol{Z}$$

其中 $\boldsymbol{Z} = (Z_1, Z_2)$是一对符合 $N(0, 1)$的独立随机变量，而 A 是一个 2×2 的矩阵．这里的核心思想是，噪声向量 $A\boldsymbol{Z}$ 的分量总体上并不独立．比如，如果 A 的两行是完全一致的，则 $A\boldsymbol{Z}$ 的两行也会完全一样．所以，这个模型允许我们考虑两个分量错误之间的相关性．这一模型也指出，分量错误之间的相关性源于它们都是共同基本噪声源的不同线性组合．

在这个模型下应当如何计算 $MLE[\boldsymbol{X}|\boldsymbol{Y}]$?

$$f_{Y|X}[\boldsymbol{y} \mid \boldsymbol{x}] = \frac{1}{2\pi \mid A \mid} \exp\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{x})'(AA')^{-1}(\boldsymbol{y}-\boldsymbol{x})\} \tag{6.10}$$

其中 A'是矩阵 A 的转置，也就是说，对于 $i, j \in \{1, 2\}$有 $A'(i, j) = A(j, i)$（下一节详细说明）．

因此，最大似然估计是使得下式最小的 $\boldsymbol{x}_k$（简单起见，假设 A 是可逆的）：

$$(\boldsymbol{y}-\boldsymbol{x})'(AA')^{-1}(\boldsymbol{y}-\boldsymbol{x}) = ||A^{-1}\boldsymbol{y}-A^{-1}\boldsymbol{x}||^2.$$

也就是说，我们希望向量 $\boldsymbol{x}_k$ 使得 $A^{-1}\boldsymbol{x}_k$ 最接近 $A^{-1}\boldsymbol{y}$.

现在看如何理解这个结果．注意

$$W := A^{-1}\boldsymbol{Y} = A^{-1}\boldsymbol{X}+\boldsymbol{Z} =: \boldsymbol{V}+\boldsymbol{Z}.$$

因此，对于给定的 $\boldsymbol{X}$ 值，$A^{-1}\boldsymbol{Y}$ 的分量是独立同分布的，满足 $N(0, 1)$. 这么一来，计算 $MLE[\boldsymbol{V}|\boldsymbol{W}=\boldsymbol{w}]$就很简单：它是在 $\boldsymbol{V}$ 的可能值集合$\{A^{-1}\boldsymbol{x}_1, \cdots, A^{-1}\boldsymbol{x}_{16}\}$中最接近 $\boldsymbol{w}$ 的值的向量. 因此，可以通过将 $\boldsymbol{V}=A^{-1}\boldsymbol{X}$ 的最大似然估计乘以 A 来获得 $\boldsymbol{X}$ 的最大似然估计：

$$MLE[\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y}] = A\times MLE[\boldsymbol{V}|\boldsymbol{W}=A^{-1}\boldsymbol{y}].$$

6.4 联合高斯分布随机变量的密度函数

在这一节里，我们会解释式（6.10）和该结果的一般形式.

首先陈述其主要定义和将要证明的结论.

联合高斯分布

定义 6.2 联合高斯分布 $N(\mu_{\boldsymbol{Y}}, \Sigma_{\boldsymbol{Y}})$随机变量

设向量 $\boldsymbol{X}$ 服从独立 $N(0, 1)$分布的随机变量. 如果随机变量 $\boldsymbol{Y}=(Y_1, \cdots, Y_n)'$满足

$$\boldsymbol{Y} = A\boldsymbol{X}+\mu_{\boldsymbol{Y}},\ \Sigma_{\boldsymbol{Y}} = AA',$$

则称其均值为 $\mu_{\boldsymbol{Y}}$、协方差为 $\Sigma_{\boldsymbol{Y}}$ 的联合高斯分布变量, 记作 $\boldsymbol{Y} =_D N(\mu_{\boldsymbol{Y}}, \Sigma_{\boldsymbol{Y}})$. ◊

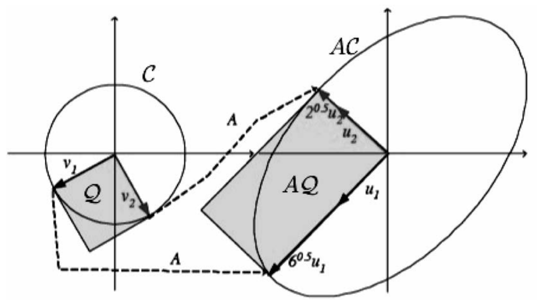


图 6-9 $f_{\boldsymbol{Y}}$的等高线

下面是主要结论.

定理 6.3 $N(\mu_{\boldsymbol{Y}}, \boldsymbol{\Sigma}_{\boldsymbol{Y}})$随机变量的密度函数

令 $\boldsymbol{Y} =_D N(\mu_{\boldsymbol{Y}}, \Sigma_{\boldsymbol{Y}})$，那么

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{\sqrt{|\Sigma_{\boldsymbol{Y}}|}(2\pi)^{\frac{n}{2}}} \exp\{-\frac{1}{2}(\boldsymbol{y}-\mu_{\boldsymbol{Y}})'\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{y}-\mu_{\boldsymbol{Y}})\}. \tag{6.11}$$

□

如图 6-9 所示，该联合概率密度函数的等高线为椭圆．附录 B.6 对其有进一步阐述．

这个联合分布是由均值和协方差矩阵所决定的．如果 $\boldsymbol{Y}' = (\boldsymbol{V}', \boldsymbol{W}')$是联合高斯分布，那么该分布由均值、$\Sigma_V$、$\Sigma_W$和 $\mathrm{cov}(\boldsymbol{V}, \boldsymbol{W})$所决定．我们知道，如果 $\boldsymbol{V}$ 和 $\boldsymbol{W}$ 相互独立，则其互不相关，即 $\mathrm{cov}(\boldsymbol{V}, \boldsymbol{W}) = 0$．由于联合分布由均值和协方差决定，可以得出结论：如果它们不相关，则它们也相互独立．我们将这一结论记为定理．

定理 6.4　联合高斯随机变量不相关，当且仅当它们是独立的

令 $\boldsymbol{V}$ 和 $\boldsymbol{W}$ 为联合高斯随机变量，那么是当且仅当它们不相关时是相互独立的. □

定理的证明会用到下面的结论．

定理 6.5　联合高斯分布的线性组合仍为联合高斯分布

令 $\boldsymbol{V}$ 和 $\boldsymbol{W}$ 为联合高斯分布变量，则 $A\boldsymbol{V}+\boldsymbol{a}$ 和 $B\boldsymbol{W}+\boldsymbol{b}$ 仍为联合高斯分布. □

证明　由定义可知，如果 $\boldsymbol{V}$ 和 $\boldsymbol{W}$ 是服从 $N(0, 1)$的独立同分布随机变量的线性函数，则 $\boldsymbol{V}$ 和 $\boldsymbol{W}$ 是联合高斯分布．那么，$A\boldsymbol{V}+\boldsymbol{a}$ 和 $B\boldsymbol{W}+\boldsymbol{b}$ 是同一组独立同分布的 $N(0, 1)$随机变量的线性函数，因此它们服从联合高斯分布．更明确地说，存在服从 $\mathcal{N}(0, 1)$的独立同分布随机变量 $\boldsymbol{X}$ 使得

$$\begin{bmatrix} \boldsymbol{V} \\ \boldsymbol{W} \end{bmatrix} = \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{d} \end{bmatrix} + \begin{bmatrix} C \\ D \end{bmatrix} \boldsymbol{X},$$

所以，

$$\begin{bmatrix} A\boldsymbol{V}+\boldsymbol{a} \\ B\boldsymbol{W}+\boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}+A\boldsymbol{c} \\ \boldsymbol{b}+B\boldsymbol{d} \end{bmatrix} + \begin{bmatrix} AC \\ BD \end{bmatrix} \boldsymbol{X}. \quad ♣$$

举一个例子，令 X 和 Y 为独立的 $N(0, 1)$随机变量．那么，$X+Y$ 和 $X-Y$ 是独立的．实际上，由定理 6.5 可得，这些随机变量是联合高斯分布且不相关的．这是因为：

$$E((X+Y)(X-Y)) - E(X+Y)E(X-Y) = E(X^2 - Y^2) = 0.$$

所以，由定理 6.4 可知它们是独立的．

本小节的剩余部分对式（6.11）进行推导．为方便起见，首先回顾一下附录 A.13 介绍的如何由 $\boldsymbol{X}$ 的概率密度计算 $A\boldsymbol{X}+\boldsymbol{b}$ 的概率密度函数：

$$f_Y(\boldsymbol{y}) = \frac{1}{|A|} f_X(\boldsymbol{x}), \text{其中 } A\boldsymbol{x}+\boldsymbol{b}=\boldsymbol{y}. \tag{6.12}$$

现在把式（6.12）应用到由 n 个 $N(0, 1)$独立随机变量所组成的向量 $\boldsymbol{X}$ 上．在这种情况下得到

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{n} f_{X_i}(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x_i^2}{2}\}$$

$$= \frac{1}{(2\pi)^{n/2}} \exp\{-\frac{\|\boldsymbol{x}\|^2}{2}\}.$$

于是，由式（6.12）得到

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{|A|} \frac{1}{(2\pi)^{n/2}} \exp\{-\frac{\|\boldsymbol{x}\|^2}{2}\},$$

其中 $A\boldsymbol{x}+\mu_{\boldsymbol{Y}} = \boldsymbol{y}$. 因此，

$$\boldsymbol{x} = A^{-1}(\boldsymbol{y}-\mu_{\boldsymbol{Y}}),$$

并且

$$\|\boldsymbol{x}\|^2 = \|A^{-1}(\boldsymbol{y}-\mu_{\boldsymbol{Y}})\|^2 = (\boldsymbol{y}-\mu_{\boldsymbol{Y}})'(AA')^{-1}(\boldsymbol{y}-\mu_{\boldsymbol{Y}}).$$

上面的公式用到了$\|\boldsymbol{z}\|^2 = \boldsymbol{z}'\boldsymbol{z}$ 和$(M\boldsymbol{v})' = \boldsymbol{v}'M'$.

现在回顾协方差矩阵的定义①：

$$\Sigma_{\boldsymbol{Y}} = E((\boldsymbol{Y}-E(\boldsymbol{Y}))(\boldsymbol{Y}-E(\boldsymbol{Y}))')$$

由于 $\boldsymbol{Y} = A\boldsymbol{X}+\mu_{\boldsymbol{Y}}$，$\Sigma_{\boldsymbol{X}} = \boldsymbol{I}$（$\boldsymbol{I}$ 为单位阵），则

$$\Sigma_{\boldsymbol{Y}} = A\Sigma_{\boldsymbol{X}}A' = AA'.$$

并且，

$$|\Sigma_{\boldsymbol{Y}}| = |A|^2.$$

综上可得

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{\sqrt{|\Sigma_{\boldsymbol{Y}}|(2\pi)^{n/2}}} \exp\{-\frac{1}{2}(\boldsymbol{y}-\mu_{\boldsymbol{Y}})'\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{y}-\mu_{\boldsymbol{Y}})\}.$$

这正是式（6.11）.

6.5 奈曼—皮尔逊定理 5.6 的证明

证明思路是考虑任意一个满足 $P[\tilde{X} = 1|X = 0] \leqslant \beta$ 的估计 $\tilde{X}$ 的决策规则，然后证明

$$P[\tilde{X} = 1|X = 1] \leqslant P[\hat{X}|X = 1]. \qquad (6.13)$$

其中，$\hat{X}$ 为定理 5.6 的估计. 为了证明该式，我们注意到

$$(\hat{X} - \tilde{X})(L(Y) - \lambda) \geqslant 0.$$

事实上，当 $L(Y)-\lambda > 0$ 时，有 $\hat{X} = 1 \geqslant \tilde{X}$，所以上式的确是非负的. 当 $L(Y)-\lambda < 0$ 时，有 $\hat{X} = 0 \leqslant \tilde{X}$，因而上式还是非负的.

① 参见附录 A2.

对该式两边取给定 $X=0$ 的条件期望，

$$E[\hat{X}L(Y)|X=0]-E[\tilde{X}L(Y)|X=0]\geqslant\lambda(E[\hat{X}|X=0]-E[\tilde{X}|X=0]).\tag{6.14}$$

根据定理，得到

$$E[\hat{X}|X=0]=P[\hat{X}=1|X=0]=\beta\geqslant P[\tilde{X}=1|X=0]=E[\tilde{X}|X=0].$$

因此，由式（6.14）得出

$$E[\hat{X}L(Y)|X=0]\geqslant E[\tilde{X}L(Y)|X=0].\tag{6.15}$$

对于任意函数 $g(Y)$ 有

$$\begin{aligned}E[g(Y)L(Y)\mid X=0]&=\int g(y)L(y)f_{Y|X}(y\mid 0)\mathrm{d}y\\&=\int g(y)\frac{f_{Y|X}[y\mid 1]}{f_{Y|X}[y\mid 0]}f_{Y|X}[y\mid 0]\mathrm{d}y\\&=\int g(y)f_{Y|X}[y\mid 1]\mathrm{d}y\\&=E[g(Y)\mid X=1].\end{aligned}$$

注意，即使将函数 $g(Y)$ 换成 $g(Y, Z)$（Z 是独立于 X 和 Y 的随机变量），这一结果仍然成立．现在取 $g(Y)=X$，我们得到

$$E[\hat{X}L(Y)|X=0]=E[\hat{X}|X=1]=P[\hat{X}=1|X=1].$$

类似地，

$$E[\tilde{X}L(Y)|X=0]=P[\tilde{X}=1|X=1].$$

将上述两式和式（6.15）相结合，即可得出式（6.13）.

6.6　小结

- LDPC 码
- 联合高斯分布随机变量，当且仅当不相关时独立
- 奈曼–皮尔逊定理的证明.

重要方程与公式

LDPC	$\boldsymbol{y}=H\boldsymbol{x}$	式（6.1）
P(奇数)	$P(\sum_j X_j=1)=0.5-0.5\prod_j(1-2p_j)$	式（6.2）
专家观点的融合	$P[X=1\mid Y_1,\cdots,Y_n]=\prod_j p_j/(\prod_j p_j+\prod_j \bar{p}_j)$	式（6.5）
联合高斯分布	$N(\mu,\Sigma)\ \Leftrightarrow f_X=\ \cdots$	式（6.10）
若$\boldsymbol{X}$和$\boldsymbol{Y}$是联合高斯分布，则	$\boldsymbol{X}\perp\boldsymbol{Y}\Rightarrow\boldsymbol{X},\boldsymbol{Y}$独立	定理6.4

6.7 参考资料

Richardson 和 Urbanke 在 *Modern Coding Theory* 一书中全面地介绍了 LDPC 码以及迭代式解码技术.

6.8 练习

1. 构建两个分别为高斯分布但不是联合高斯分布的随机变量.

提示：令 $X =_D \mathcal{N}(0, 1)$，Z 为独立的随机变量且 $P(Z = 1) = P(Z = -1) = 1/2$．定义 $Y = XZ$．证明 X 和 Y 满足题目要求.

2. 假设 $X =_D (Y+Z)/\sqrt{2}$，其中 Y 和 Z 独立同分布．证明 $X = \mathcal{N}(0, \sigma^2)$，$\sigma^2 \geqslant 0$.

提示：首先证明 $E(X) = 0$，其次递归证明 $X =_D (V_1+\cdots+V_m)/\sqrt{m}$，$m = 2^n$，且 V_m 是分布如 X 的独立同分布随机变量．运用中心极限定理得出结论.

3. 假设 $\boldsymbol{Z} =_D \mathcal{N}(\boldsymbol{0}, \Sigma)$，且

C.6

$$\Sigma = \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.3 \end{bmatrix}.$$

假设字符的先验概率相同且接收机采用最大后验机制进行解码. 用 Matlab 模拟该系统来估计出错率.

第7章

追踪定位—A

应用：估计，追踪定位

主题：LLSE，MMSE，卡尔曼滤波器

背景知识：附录 A

雷达可以发射电磁波并且接收物体反射回来的电磁波，因此被广泛应用于物体定位．GPS 接收器可以通过接收卫星发射的信号，来判断自己的位置．电脑可以通过接收温度和压力传感器的数据来判断化学反应器内的状态．同理，汽车中的控制电脑会接收不同传感器的指标来判断汽车的状态．本章主要讨论在无法观测到全部信息的情况下，如何通过有限的观测量来估计一个随机变量或随机向量．



图 7-1　用雷达信号估计物体的位置



图 7-2　用卫星信号定位设备

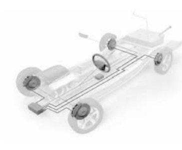


图 7-3　用传感器信号估计汽车的状态

7.1 估计问题

基本的估计问题可以用如下方法描述. 设有两个连续的随机变量(X, Y). 我们希望根据 Y 的观测值来为 X 估值.

这个问题有以下几个规范化的形式.

- 已知分布：事先知道(X, Y)的联合分布.
- 离线：事先观测一组(X, Y)的样本值.
- 在线：持续地观测(X, Y)的样本值.

我们要寻找一个推理函数 $g(\cdot)$，使得期望误差 $C(g)$最小，其中

$$C(g) = E(c(X, g(Y))) .$$

在上述表达式中，$c(X, \hat{X})$ 表示实际值为 X 的时候，猜测 $\hat{X}$ 值的代价. 一个常用的 $c(X,\ \hat{X})$函数为：

$$c(X, \hat{X}) = |X - \hat{X}|^2 .$$

我们还会研究 X 在高维度里的情况，即 $X \in \Re^d$ ，$d > 1$. 在这种情况下，可以采用$c(X, \hat{X}) = \|X - \hat{X}\|^2$. 此时，相应的最佳估值被称为已知 Y 条件下 X 的 LSE （Least Square Estimate，最小二乘估计）. 如果函数 $g(\cdot)$是任意的，则称其为已知 Y 条件下 X 的 MMSE（Minimum Mean Square Estimate，最小均方估计）. 如果函数 $g(\cdot)$是线性的，即具有 $a+bY$ 的形式，则称其为已知 Y 条件下 X 的 LLSE（Linear Least Square Estimate，线性最小平方估计）. 还可以把 $g(\cdot)$约束成一个给定阶数的多项式函数. 例如，可以定义已知 Y 条件下 X 的 QLSE（Quadratic Least Square Estimate，四次最小方差估计）. 图 7-4 绘出了这些函数.

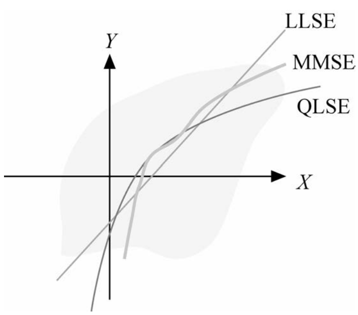


图 7-4 已知 Y 条件下 X 的最小平方估计：LLSE 是线性的，
QLSE 是四次方的，MMSE 则可以是任意的函数

在离线推理的问题中，一个常用的方法是选择一组参数化的函数 $\{g_w, w \in \Re^d\}$ ，

使如下的经验误差最小：

$$\sum_{k=1}^{K} c(X_k, g_w(Y_k)),$$

其中 w 是需要优化的参数，(X_k, Y_k) 是观测到的离散点.

对于在线问题，我们同样选取一组参数化的函数，然后采取如下的随机梯度算法：

$$w(k+1) = w(k) - \gamma \nabla_w c(X_{k+1}, g_w(Y_{k+1})),$$

其中 ∇ 是关于 w 的梯度，$\gamma > 0$ 是一个很小的步长. 这种方法背后的原理是，由于 γ 很小，根据强大数定律，参数的更新会趋向于

$$-\sum_{i=k}^{k+K-1} \nabla_w c(X_{i+1}, g_w(Y_{i+1})) \approx -K\nabla E(c(X_k, g_w(Y_k))) = -K\nabla C(g_w).$$

熟悉梯度算法的读者可以看到，这一项对应的是能够最小化 $C(g_w)$ 的梯度算法.

7.2　线性最小平方估计（LLSE）

学习 LLSE 之前，回顾一下前面一节中提出的设定：有一对已知联合分布的随机变量(X, Y)，要寻找函数 $g(Y) = a+bY$ 来最小化下面的式子：

$$C(g) = E(|X - g(Y)|^2).$$

考虑三种情况：已知随机变量的分布；已有一系列观测值；每一次能够观测到一个新样本点的情况.

假设已知联合分布(X, Y). 这意味着我们知道其联合累积分布函数 $F_{X,Y}(x, y)$[①].

寻找函数 $g(Y) = a+bY$ 来最小化下面的式子：

$$C(g) = E(|X - g(Y)|^2) = E(|X - a - bY|^2).$$

我们用 $L(X|Y)$ 来表示这个函数. 下面是 $L(X|Y)$ 的正式定义.

定义 7.1　LLSE

已知 Y 条件下 X 的 LLSE $L(X|Y)$ 是具有 $a+bY$ 形式的函数，能够最小化

$$E(|X - a - bY|^2). \qquad \Diamond$$

注意到

$$\begin{aligned} C(g) &= E(X^2 + a^2 + b^2Y^2 - 2aX - 2bXY + 2abY) \\ &= E(X^2) + a^2 + b^2E(Y^2) - 2aE(X) - 2bE(XY) - 2abE(Y). \end{aligned}$$

为了找到能够使这个表达式最小的 a 和 b 值，将上式对 a 和 b 分别求偏导，并令偏导

① 参见附录 A.

为 0．由此得到

$$0 = 2a - 2E(X) + 2bE(Y), \tag{7.1}$$

$$0 = 2bE(Y^2) - 2E(XY) + 2aE(Y). \tag{7.2}$$

通过求解这两个关于 a 和 b 的方程，得到

$$L(X|Y) = a + bY = E(X) + \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(Y)}(Y - E(Y)).$$

上式用到了下面的等式：

$$\mathrm{cov}(X,Y) = E(XY) - E(X)E(Y) \text{ 以及 } \mathrm{var}(Y) = E(Y^2) - E(Y)^2.$$

将其总结成下面的定理.

定理 7.2　LLSE

$$L(X|Y) = E(X) + \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(Y)}(Y - E(Y)). \tag{7.3}$$

□

举一个例子．设

$$Y = \alpha X + Z, \tag{7.4}$$

其中 X 和 Z 是两个均值为 0 的独立随机变量．由独立性可以得到①

$$\begin{aligned}\mathrm{cov}(X,Y) &= E(XY) - E(X)E(Y)\\ &= E(X(\alpha X + Z)) = \alpha E(X^2)\end{aligned}$$

$$\mathrm{var}(Y) = \alpha^2\,\mathrm{var}(X) + \mathrm{var}(Z) = \alpha^2 E(X^2) + E(Z^2).$$

因此，

$$L(X|Y) = \frac{\alpha E(X^2)}{\alpha^2 E(X^2) + E(Z^2)} Y = \frac{\alpha^{-1} Y}{1 + SNR^{-1}},$$

其中，

$$SNR := \frac{\alpha^2 E(X^2)}{\sigma^2}$$

被称为信噪比，即 Y 的信号功率 $E(\alpha^2 X^2)$ 除以噪声功率 $E(Z^2)$．注意，如果 SNR 很小，那么 $L(X|Y)$ 接近于 0. 这是在没有任何观测量的情况下对 X 的最好估计值. 如果 SNR 非常大，那么 $L(X|Y) \approx \alpha^{-1} Y$，这也是当 $Z = 0$ 时的正确猜测值.

再举一例．设

$$X = \alpha Y + \beta Y^2, \tag{7.5}$$

① $E(XZ) = E(X)E(Z) = 0$

其中[①] $Y =_D U[0,1]$，则

$$
\begin{aligned}
E(X) &= \alpha E(Y) + \beta E(Y^2) = \alpha/2 + \beta/3; \\
\mathrm{cov}(X,Y) &= E(XY) - E(X)E(Y) \\
&= E(\alpha Y^2 + \beta Y^3) - (\alpha/2 + \beta/3)(1/2) \\
&= \alpha/3 + \beta/4 - \alpha/4 - \beta/6 \\
&= (\alpha+\beta)/12 \\
\mathrm{var}(Y) &= E(Y^2) - E(Y^2) = 1/3 - (1/2)^2 = 1/12.
\end{aligned}
$$

因此，

$$L(X \mid Y) = \alpha/2 + \beta/3 + (\alpha+\beta)(Y - 1/2) = -\beta/6 + (\alpha+\beta)Y.$$

图 7-5 绘出了这个估计值．显然，通过观测 Y 可以计算 X．不过，这里只考虑线性的 $L(X \mid Y)$ 函数．

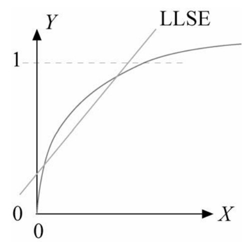


图 7-5　当 $Y =_D U[0,1]$ 时 $L(\alpha Y + \beta Y^2 \mid Y)$ 的值

投影

从本质上来说，$L(X \mid Y)$ 可以被视为一种投影．这种思想对于后面理解更加复杂的估计问题很有帮助．具体来说，把 $L(X \mid Y)$ 看成 X 到（由 Y 的线性函数组成的）函数空间 $\mathcal{L}(Y)$的投影．

图 7-6 绘出了这种诠释方法．在图中，随机变量由点来表示，而 $\mathcal{L}(Y)$构成了平面，这是因为该集合中点的线性组合仍然属于这个集合（线性函数的线性组合仍为线性函数）．在图中，从一个随机变量 V 到另一个随机变量 W 的模长平方为 $E(|V-W|^2)$．同时，当 $E(VW) = 0$ 时，向量 V 和 W 相互正交．因此，如果 $L(X \mid Y) = a + bY$ 是 X 到 $\mathcal{L}(Y)$ 的投影，则 $X - L(X \mid Y)$ 与每一个关于 Y 的线性函数正交，即

$$E((X - a - bY)(c + dY)) = 0, \forall c, d \in \Re .$$

---

① 因此，$E(Y^k) = (1+k)^{-1}$．

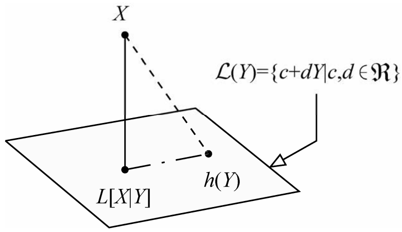


图 7-6　$L(X\mid Y)$ 是 X 到 $L(Y)$的投影

另一种等价的说法是，

$$E(X)=a+bE(Y)\text{，并且 }E((X-a-bY)Y)=0\,. \qquad (7.6)$$

这两个式子与式（7.1）和式（7.2）是一样的．我们将式（7.6）称为投影性质．

图 7-7 阐述了如下情况的投影：

$$X=\mathcal{N}(0,1)\text{，}Y=X+Z\text{，其中 }Z=\mathcal{N}(0,\sigma^2)\,.$$

在图中，Z 的模长等于 $\sqrt{E(Z^2)}=\sigma$，X 的模长为 $\sqrt{E(X^2)}=1$，且向量 X 与 Z 相互正交（因为 $E(XZ)=0$）.

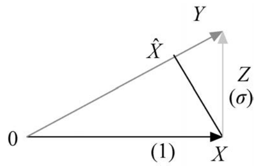


图 7-7　投影的例子

可以看出三角形 $0\hat{X}X$ 与 $0XY$ 是相似的．因此，

$$\frac{\|\hat{X}\|}{\|X\|}=\frac{\|X\|}{\|Y\|'}\,.$$

因为 $\|Y\|=\sqrt{1+\sigma^2}$，所以，

$$\frac{\|\hat{X}\|}{1}=\frac{1}{\sqrt{1+\sigma^2}}=\frac{\|Y\|}{1+\sigma^{2\prime}}\,.$$

这说明

$$\hat{X}=\frac{1}{1+\sigma^2}Y\,.$$

在图 7-6 中可以看到，投影性质保证了 $L[X\mid Y]$ 是 $\mathcal{L}(Y)$中离 X 最近的点．事实上，这一点可以通过验证下式得到：

$$E(|X-L[X\mid Y]|^2)\leqslant E(|X-h(Y)|^2)$$

对于所有的 $h(Y)=c+dY$ 均成立．证明这一不等式的思路是在以 X、$L[X\mid Y]$ 和 $h(Y)$为顶点的三角形中应用勾股定理：

$$\begin{aligned}E(|X-h(Y)|^2) &= E(|X-L[X\mid Y]+L[X\mid Y]-h(Y)|^2)\\ &= E(|X-L[X\mid Y]|^2)+E(|L[X\mid Y]-h(Y)|^2)\\ &\quad +2E((X-L[X\mid Y])(L[X\mid Y]-h(Y))).\end{aligned}$$

事实上，由于 $L[X\mid Y]-h(Y)$ 是关于 Y 的线性函数，投影性质式（7.6）说明了上式的最后一项等于 0．于是，

$$\begin{aligned}E(|X-h(Y)|^2) &= E(|X-L[X\mid Y]|^2)+E(|L[X\mid Y]-h(Y)|^2)\\ &\geqslant E(|X-L[X\mid Y]|^2),\end{aligned}$$

结论得证．

7.3 线性回归

现在假设我们并不知道(X, Y)的联合分布，而是观测到了 K 个独立同分布的样本$(X_1,Y_1),\cdots,(X_K,Y_K)$．希望寻找一个函数 $g(Y)=a+bY$ 使得

$$E(|X-a-bY|^2)$$

最小．我们采取的方法是通过选取 a 和 b 的值，让基于这些样本点的误差平方和最小．也就是说，通过选取 a 和 b 的值使得下式最小：

$$\sum_{k=1}^{K}|X_k-a-bY_k|^2 .$$

为此，将上式分别对 a 和 b 求偏导，并令偏导为 0，可知 a 和 b 满足下式：

$$a+bY=E_K(X)+\frac{\operatorname{cov}_K(X,Y)}{\operatorname{var}_K(Y)}(Y-E_K(Y)) . \tag{7.7}$$

这里定义

$$E_K(X)=\frac{1}{K}\sum_{k=1}^{K}X_k,\ E_K(Y)=\frac{1}{K}\sum_{k=1}^{K}Y_k,$$

$$\operatorname{cov}_K(X,Y)=\frac{1}{K}\sum_{k=1}^{K}X_kY_k-E_K(X)E_K(Y),$$

$$\operatorname{var}_K(Y)=\frac{1}{K}\sum_{k=1}^{K}Y_k^2-E_K(Y)^2.$$

也就是说，式（7.7）与式（7.3）基本相同，唯一的区别是由样本均值取代了期望．式（7.7）被称为由 Y 到 X的线性回归，如图 7-8 所示．

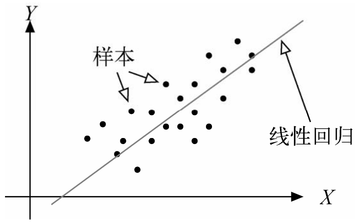


图 7-8　由 Y 到 X 的线性回归

下面的定理说明了线性回归与 LLSE 的关系.

定理 7.3　线性回归收敛于 LLSE

随着样本点的增加，线性回归的结果将收敛于 LLSE. □

证明　当 $K\to\infty$ 时，根据强大数定律，

$$E_K(X)\to E(X), E_K(Y)\to E(Y),$$
$$\operatorname{cov}_K(X,Y)\to\operatorname{cov}(X,Y), \operatorname{var}_K(Y)\to\operatorname{var}(Y).$$

结合线性回归的表达式以及 LLSE 的结论就可以得出结论. ♣

式（7.3）以及线性回归为协方差提供了一种更直观的理解. 如果协方差为 0，则 $L[X|Y]$ 与 Y 不相关. 如果协方差为正（负），则它随着 Y 而增大（减小）. 因此，$\operatorname{cov}(X,Y)$ 可以被视为线性回归意义下对相关性的一种度量. 例如，图 7-9 中的随机变量是彼此不相关的，因为 $L[X|Y]$ 并不依赖于 Y.

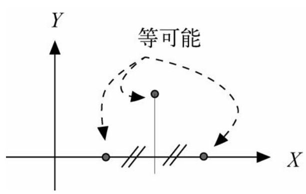


图 7-9　随机变量 X 与 Y 彼此不相关，但它们并不是相互独立的

7.4　最小均方估计（MMSE）

上一节研究了通过寻找线性方程 $a+bY$ 来近似 X 使得最小平方误差最小的问题. 事实上，也可以将结论推广到二次方近似 $a+bY+cY^2$ 或者任意阶的多项式形式近似. 其基本思想是相同的，并且同样可以通过投影的概念来理解.

原则上，高阶的多项式函数可以更好地拟合 X，因为相对于低阶的多项式函数，它们有更多的多项式函数可供选择. 但是这时候，让全部参数符合给定的观测量会更

为复杂．例如，如果有 K 个样本点，就可以找到一个 K 阶的多项式函数来准确地拟合这些观测点．可是，这并不意味着采用这个多项式函数就会比采用低阶的多项式函数得到更小的均方误差．这种情况叫作过度拟合．

现在，假设知道 (X, Y) 的联合分布函数，希望寻找一个函数 $g(Y)$ 使得 $E(|X-g(Y)|^2)$ 最小．最好的函数就是给定 Y 条件下 X 的 MMSE．有如下定理．

定理 7.4　MMSE 是一个条件期望

给定 Y 条件下 X 的 MMSE 为

$$g(Y) = E[X|Y],$$

其中 $E[X|Y]$ 是给定 Y 条件下 X 的**条件期望**. □

在证明这一结论之前，首先定义条件期望．

定义 7.5　条件期望

给定 Y 条件下 X 的条件期望定义为

$$E[X \mid Y=y] = \int_{-\infty}^{\infty} x f_{X|Y}[x \mid y]\,\mathrm{d}x,$$

其中

$$f_{X|Y}[x \mid y] := \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

是给定 Y 条件下 X 的条件概率密度函数． ◊

图 7-10 阐述了条件期望的含义．图中假定均匀地从阴影区域中选取 (X, Y)．因此，如果观测到 $Y \in (y, y+\mathrm{d}y)$，那么 X 点均匀分布在 $Y = y$ 的长条阴影区域中．因此，X 的平均值是这个长条区域的中点，如图所示．图中的点划线展示了 Y 如何决定均值以及 $E[X|Y]$．

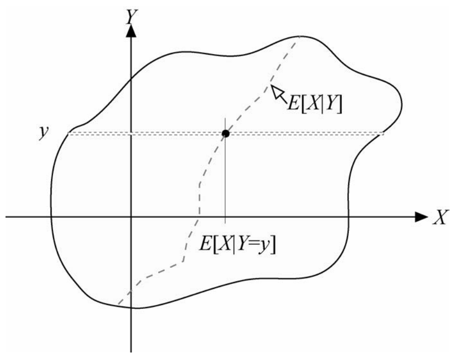


图 7-10　从阴影区域中均匀地选取 (X, Y) 时，条件期望 $E[X|Y]$ 的情况

下面的结论可直接由定义得到.

引理 7.6 MMSE 的正交性质

(a) 对于任意函数 $\phi(\cdot)$，有

$$E((X - E[X \mid Y])\phi(Y)) = 0. \qquad (7.8)$$

(b) 如果函数 $g(Y)$满足

$$E((X - g(Y))\phi(Y)) = 0, \forall \phi(\cdot), \qquad (7.9)$$

则 $g(Y) = E[X \mid Y]$.

证明

(a) 为证明式（7.8），注意

$$\begin{aligned} E(E[X \mid Y]\phi(Y)) &= \int_{-\infty}^{\infty} E[X \mid Y = y]\phi(y) f_Y(y)\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \frac{f_{X,Y}(x,y)}{f_Y(y)} \mathrm{d}x \phi(y) f_Y(y)\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\phi(y) f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y \\ &= E(X\phi(Y)). \end{aligned}$$

式（7.8）得证.

(b) 为证明引理第二部分，注意让式（7.9）中 $\phi(Y) = g(Y) - E[X \mid Y]$，并使用式（7.8）. 我们得到：

$$\begin{aligned} &E(| g(Y) - E[X \mid Y]|)^2 \\ &\quad = E((g(Y) - E[X \mid Y])\{(g(Y) - X) - E([X \mid Y] - X)\}) = 0, \end{aligned}$$

注意到，引理的第二部分说明了：投影性质唯一地决定了条件期望．也就是说，X 到 $\mathcal{G}(Y)$的投影只有一个. ♣

现在来证明定理 7.4.

定理 7.4 的证明

等式（7.8）是投影性质．它说明了 $X - E[X \mid Y]$ 垂直于 Y 的函数族 $\mathcal{G}(Y)$，如图 7-11 所示．因此，它也垂直于 $h(Y) - E[X \mid Y]$．和 LLSE 的情形相似，这个投影性质说明对任意函数 $h(\cdot)$都有

$$E(| X - h(Y) |^2) \geqslant E(| X - E[X \mid Y]|^2).$$

这就说明 $E[X \mid Y]$ 确实是给定 Y 条件下 X 的 MMSE. ♣

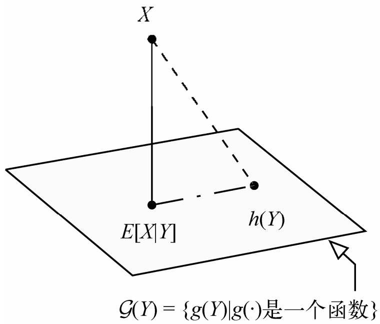


图 7-11　条件期望 $E[X\mid Y]$ 就是 X 到 Y 的函数族 $\mathcal{G}(Y)$的投影

根据定义,我们知道如何通过给定 Y 条件下 X 的条件概率分布来求得 $E[X\mid Y]$.然而，很多情况下可以通过更简单的方法求得 $E[X\mid Y]$. 下面的方法利用了条件期望的性质.

定理 7.7　条件期望的性质

(a) 线性：$E[a_1X_1+a_2X_2\mid Y]=a_1E[X_1\mid Y]+a_2E[X_2\mid Y]$

(b) 分解已知量：$E[h(Y)X\mid Y]=h(Y)E[X\mid Y]$

(c) 光滑性：$E(E[X\mid Y])=E(X)$

(d) 独立性：当 X 与 Y 相互独立时，$E[X\mid Y]=E(X)$　□

证明

(a) 首先，$a_i(X_i-E[X_i\mid Y])$，（$i=1,2$）垂直于 $\mathcal{G}(Y)$. 因此，

$$a_1X_1+a_2X_2-(a_1E[X_1\mid Y]+a_2E[X_2\mid Y])$$

垂直于 $\mathcal{G}(Y)$. 现在使用引理 7.6(b)即可得到结论.

(b) 只需要证明

$$h(Y)X-h(Y)E[X\mid Y]$$

垂直于 $\mathcal{G}(Y)$，即

$$E((h(Y)X-h(Y)E[X\mid Y])\phi(Y))=0, \forall\phi(.) .$$

然后结论可以由引理 7.6(b)得到. 现在，由于 $X-E[X\mid Y]$ 垂直于 $\mathcal{G}(Y)$，它也垂直于 $h(Y)\phi(Y)$. 由此，

$$E((h(Y)X-h(Y)E[X\mid Y])\phi(Y))=E((X-E[X\mid Y])h(Y)\phi(Y))=0 .$$

(c) 设式（7.8）中的 $\phi(Y)=1$ ，则

$$E(X-E[X\mid Y])=0 .$$

命题得证.

(d) 首先，

$$E((X-E(X))\phi(Y)) = E(X-E(X))E(\phi(Y)) = 0\,.$$

这里第一个等式成立的原因是 $X-E(X)$ 与 $\phi(Y)$ 相互独立（两个相互独立的随机变量，其函数也相互独立）[1]．这说明

$$X-E(X)$$

垂直于 $\mathcal{G}(Y)$．根据引理 7.6(b)，可知命题得证. ♣

下面通过一个例子来看这些结论. 假设 X、Y、Z 为独立同分布于 $U[0, 1]$的随机变量. 希望计算

$$E[(X+2Y)^2 \mid Y].$$

由于 $X =_D U[0, 1]$,

$$\begin{aligned}
E[(X+2Y)^2|Y] &= E[X^2+4Y^2+4XY|Y] \\
&= E[X^2|Y]+4E[Y^2|Y]+4E[XY|Y], \text{由期望的线性性质得到} \\
&= E(X^2)+4E[Y^2|Y]+4E[XY|Y], \text{由变量相互独立性质得到} \\
&= E(X^2)+4Y^2+4YE[X|Y], \text{在条件期望中将已知值提出期望外} \\
&= E(X^2)+4Y^2+4YE(X), \text{由变量独立性质得到} \\
&= \frac{1}{3}+4Y^2+2Y.
\end{aligned}$$

注意，如果没有以上条件期望的性质，给定 Y 时计算 $(X+2Y)^2$ 会很繁琐.

在有些情况下，还可以利用对称性来计算条件期望. 设 X、Y、Z 是独立同分布的，那么

$$E[X \mid X+Y+Z] = \frac{1}{3}(X+Y+Z)\,. \tag{7.10}$$

式（7.10）利用了对称性，

$$E[X \mid X+Y+Z] = E[Y \mid X+Y+Z] = E[Z \mid X+Y+Z].$$

设它们的值都是 V．根据线性，它们的和为

$$3V = E[X+Y+Z \mid X+Y+Z].$$

因此，$3V = X+Y+Z$，结论得证.

联合高斯分布的MMSE

一般来说，$L[X \mid Y] \neq E[X \mid Y]$．举一个简单的例子，设 $Y =_D U[-1,1]$ 且 $X = Y^2$，则 $\text{cov}(X,Y) = E(XY) - E(X)E(Y) = 0$．因此，$E[X \mid Y] = Y^2$，但 $L[X \mid Y] = E(X) = 1/3$．

---

① 参见附录 A.

图 7-12 强调了 $E[X|Y]$ 是 X 到 $\mathcal{G}(Y)$ 的投影，$L[X|Y]$ 是 X 到 $\mathcal{L}(Y)$ 的投影．因为 $\mathcal{L}(Y)$ 是 $\mathcal{G}(Y)$ 的子空间，所以这两个投影通常是不相同的．不过，有些时候 $E[X|Y]$ 碰巧是线性的，比如式（7.10）．当然，也不难构造更多其他的例子．

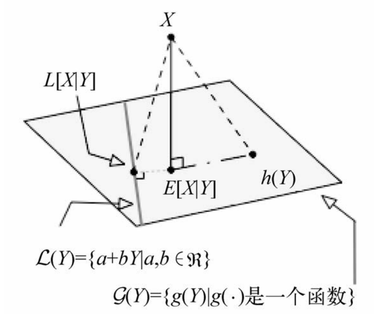


图 7-12　通常来说，MMSE 和 LLSE 的结果是不同的

不过，这一结果在当 X 和 Y 为联合高斯分布的情况下也成立．由于联合高斯问题的重要性，我们将这一结论单独总结成如下定理．

定理 7.8　联合高斯分布的 MMSE

设 X 和 Y 是服从联合高斯分布的随机变量，则

$$E[X|Y]=L[X|Y]=E(X)+\frac{\operatorname{cov}(X,Y)}{\operatorname{var}(Y)}(Y-E(Y)). \quad \square$$

证明

首先，我们注意到

$$X-L[X|Y] \text{ 与 Y 是不相关的.}$$

同时，$X-L[X|Y]$ 与 Y 是服从联合高斯分布的随机变量 X 和 Y 的两个线性函数．所以，由定理 6.5 可知它们服从联合高斯分布．并且由定理 6.4 可知它们相互独立．

因此，对任意 $\phi(\cdot)$，

$$X-L[X|Y] \text{ 与 } \phi(Y) \text{ 相互独立.}$$

这是因为彼此独立的随机变量的函数也彼此独立（这一性质在附录 A 的定理 A.9 中有说明）．于是，由附录 A 的定理 A.6 可知，对所有的 $\phi(\cdot)$，

$$X-L[X|Y] \text{ 与 } \phi(Y) \text{ 不相关.}$$

这表明

$$X-L[X|Y] \text{ 垂直于 } \mathcal{G}(Y),$$

因此，$L[X|Y] = E[X|Y]$. ♣

7.5 随机向量的情况

到目前为止,我们把关注点放在了讨论 X 和 Y 是标量随机变量情况下的 $L[X\,|\,Y]$ 和 $E[X\,|\,Y]$. 在这一节,我们将考虑随机向量的情况,即 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 是随机向量时的 $L[\boldsymbol{X}\,|\,\boldsymbol{Y}]$ 和 $E[\boldsymbol{X}\,|\,\boldsymbol{Y}]$. 这里唯一的难点在于数学符号的变化,其余的概念与思想和前面一样.

定义 7.9 随机向量的 LLSE

设 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 分别是 m 和 n 维的随机向量. 则

$$L[\boldsymbol{X}\,|\,\boldsymbol{Y}] = A\boldsymbol{Y} + \boldsymbol{b},$$

其中 A 是 $m\times n$ 的矩阵,$\boldsymbol{b}$ 是 $\Re^m$ 的向量,使得

$$E(\|\boldsymbol{X} - A\boldsymbol{Y} - \boldsymbol{b}\|^2) \text{ 最小}.$$ ◊

与标量的情况一样,LLSE 是在最小均方误差意义下对 $\boldsymbol{X}$ 的最好的线性估值.

下面的推导会用到 A.5 节中关于 $\sum_{\boldsymbol{Y}}$ 和 $\mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y})$的概念.

定理 7.10 随机向量的 LLSE

设 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 都是随机向量,且 $\sum_{\boldsymbol{Y}}$ 是满秩的,

(a) 则,

$$L[\boldsymbol{X}\,|\,\boldsymbol{Y}] = E(\boldsymbol{X}) + \mathrm{cov}(\boldsymbol{X},\boldsymbol{Y})\textstyle\sum_{\boldsymbol{Y}}^{-1}(\boldsymbol{Y} - E(\boldsymbol{Y})). \tag{7.11}$$

(b) 更进一步,

$$E(\|\boldsymbol{X} - L[\boldsymbol{X}\,|\,\boldsymbol{Y}]\|^2) = \mathrm{tr}(\textstyle\sum_{\boldsymbol{X}} - \mathrm{cov}(\boldsymbol{X},\boldsymbol{Y})\sum_{\boldsymbol{Y}}^{-1}\mathrm{cov}(\boldsymbol{Y},\boldsymbol{X})). \tag{7.12}$$

在这个表达式中,对于方阵 M,$\mathrm{tr}(M) := \sum_i M_{i,i}$ 为矩阵的迹. □

证明

(a) 对向量的证明与标量的情况类似. 设 $\boldsymbol{Z}$ 为式(7.11)的右边. 下面先推出误差 $\boldsymbol{X}-\boldsymbol{Z}$ 与所有的 $\boldsymbol{Y}$ 的线性函数正交. 然后由这一点推出 $\boldsymbol{X}$ 比所有的 $\boldsymbol{Y}$ 的线性函数 $h(\boldsymbol{Y})$ 都更靠近 $\boldsymbol{Z}$.

首先来证明正交性. 因为 $E(\boldsymbol{X}-\boldsymbol{Z}) = 0$,所以

$$E((\boldsymbol{X} - \boldsymbol{Z})(B\boldsymbol{Y} + \boldsymbol{b})') = E((\boldsymbol{X} - \boldsymbol{Z})(B\boldsymbol{Y})') = E((\boldsymbol{X} - \boldsymbol{Z})\boldsymbol{Y}')B'.$$

然后证明 $E((\boldsymbol{X} - \boldsymbol{Z})\boldsymbol{Y}') = 0$. 注意到

$$
\begin{aligned}
E((\boldsymbol{X}-\boldsymbol{Z})\boldsymbol{Y}') &= E((\boldsymbol{X}-\boldsymbol{Z})(\boldsymbol{Y}-E(\boldsymbol{Y}))') \\
&= E((\boldsymbol{X}-E(\boldsymbol{X}))(\boldsymbol{Y}-E(\boldsymbol{Y}))') \\
&\quad -\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})\Sigma_{\boldsymbol{Y}}^{-1}E((\boldsymbol{Y}-E(\boldsymbol{Y}))(\boldsymbol{Y}-E(\boldsymbol{Y}))') \\
&= \operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})-\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})\Sigma_{\boldsymbol{Y}}^{-1}\Sigma_{\boldsymbol{Y}}=0.
\end{aligned}
$$

接下来，证明 $\boldsymbol{Z}$ 比 $\boldsymbol{Y}$ 的所有线性函数 $h(\boldsymbol{Y})$都更靠近 $\boldsymbol{X}$.

$$
\begin{aligned}
E(\|\boldsymbol{X}-h(\boldsymbol{Y})\|^2) &= E((\boldsymbol{X}-h(\boldsymbol{Y}))'(\boldsymbol{X}-h(\boldsymbol{Y}))) \\
&= E((\boldsymbol{X}-\boldsymbol{Z}+\boldsymbol{Z}-h(\boldsymbol{Y}))'(\boldsymbol{X}-\boldsymbol{Z}+\boldsymbol{Z}-h(\boldsymbol{Y}))) \\
&= E(\|\boldsymbol{X}-\boldsymbol{Z}\|^2)+E(\|\boldsymbol{Z}-h(\boldsymbol{Y})\|^2)+2E((\boldsymbol{X}-\boldsymbol{Z})'(\boldsymbol{Z}-h(\boldsymbol{Y}))).
\end{aligned}
$$

现在证明最后一项等于 0．注意到

$$
E((\boldsymbol{X}-\boldsymbol{Z})'(\boldsymbol{Z}-h(\boldsymbol{Y})))=\sum_{i=1}^{n}E((X_i-Z_i)(Z_i-h_i(\boldsymbol{Y}))).
$$

同时

$$
E((X_i-Z_i)(Z_i-h_i(\boldsymbol{Y})))=E((\boldsymbol{X}-\boldsymbol{Z})(\boldsymbol{Z}-h(\boldsymbol{Y}))')_{i,i}.
$$

因为 $\boldsymbol{X}$–$\boldsymbol{Z}$ 垂直于所有 $\boldsymbol{Y}$ 的线性函数，包括 $\boldsymbol{Z}$–$h(\boldsymbol{Y})$，所以矩阵 $E((\boldsymbol{X}-\boldsymbol{Z})(\boldsymbol{Z}-h(\boldsymbol{Y}))')$ 等于 0.

（注意：证明最后一项为 0 的另一种方法是

$$
E((\boldsymbol{X}-\boldsymbol{Z})'(\boldsymbol{Z}-h(\boldsymbol{Y})))=\operatorname{tr}E((\boldsymbol{X}-\boldsymbol{Z})(\boldsymbol{Z}-h(\boldsymbol{Y}))')=0,
$$

其中，第一个等式成立是因为 $\operatorname{tr}(AB)=\operatorname{tr}(BA)$对于维数相容（即适合运算）的矩阵都成立.）

(b) 设 $\tilde{\boldsymbol{X}}:=\boldsymbol{X}-\boldsymbol{L}[\boldsymbol{X}\mid\boldsymbol{Y}]$ 为估计的误差，则

$$
\tilde{\boldsymbol{X}}:=\boldsymbol{X}-E[\boldsymbol{X}]-\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{Y}-E(\boldsymbol{Y})).
$$

设 $\boldsymbol{V}$ 和 $\boldsymbol{W}$ 是两个均值为 0 的随机向量，M 是一个矩阵，则

$$
\begin{aligned}
\operatorname{cov}(\boldsymbol{V}-M\boldsymbol{W}) &= E((\boldsymbol{V}-M\boldsymbol{W})(\boldsymbol{V}-M\boldsymbol{W})') \\
&= E(\boldsymbol{V}\boldsymbol{V}'-2M\boldsymbol{W}\boldsymbol{V}'+M\boldsymbol{W}\boldsymbol{W}'M') \\
&= \operatorname{cov}(\boldsymbol{V})-2M\operatorname{cov}(\boldsymbol{W},\boldsymbol{V})+M\operatorname{cov}(\boldsymbol{W})M'.
\end{aligned}
$$

因此，

$$
\begin{aligned}
\operatorname{cov}(\tilde{\boldsymbol{X}}) &= \Sigma_{\boldsymbol{X}}-2\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})\Sigma_{\boldsymbol{Y}}^{-1}\operatorname{cov}(\boldsymbol{Y},\boldsymbol{X}) \\
&\quad +\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})\Sigma_{\boldsymbol{Y}}^{-1}\Sigma_{\boldsymbol{X}}\Sigma_{\boldsymbol{Y}}^{-1}\operatorname{cov}(\boldsymbol{Y},\boldsymbol{X}) \\
&= \Sigma_{\boldsymbol{X}}-\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})\Sigma_{\boldsymbol{Y}}^{-1}\operatorname{cov}(\boldsymbol{Y},\boldsymbol{X}).
\end{aligned}
$$

最后，我们用下面的性质来完成证明。对于均值为 0 的随机向量 $\boldsymbol{V}$，

$$
E(\|\boldsymbol{V}\|^2)=E(\operatorname{tr}(\boldsymbol{V}\boldsymbol{V}'))=\operatorname{tr}(E(\boldsymbol{V}\boldsymbol{V}'))=\operatorname{tr}(\Sigma_{\boldsymbol{V}}).
$$

♣

7.6 卡尔曼滤波器

C.7

卡尔曼滤波器是一种使用系统的输出来对系统当前状态估计进行更新的算法，如图 7-13 所示．在时刻 $n(n=0, 1, \cdots)$，系统的状态记为 $X(n)$，输出记为 $Y(n)$．这些变量的演变由一组线性方程来定义：

$$X(n+1)=AX(n)+V(n), n\geqslant 0; \tag{7.13}$$

$$Y(n)=CX(n)+W(n), n\geqslant 0. \tag{7.14}$$

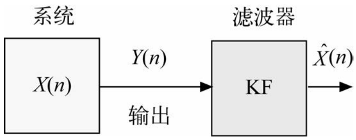


图 7-13　卡尔曼滤波器（KF）用系统上某一时刻的输出来计算系统此刻状态的 LLSE

在这组方程中，随机变量 $\{X(0), V(n), W(n), n\geqslant 0\}$ 互相正交，且均值都为 0．$V(n)$ 的协方差为 Σ_V，$W(n)$ 的协方差为 Σ_W．描述这个滤波器时，变量是随机向量，A 和 C 是维数相容的矩阵.

我们要推导出递归的等式来计算:

$$\hat{X}(n)=L[X(n)\mid Y(0),\cdots,Y(n)], n\geqslant 0.$$

7.6.1 滤波器

下面给出结论，具体的证明会在下一节给出．虽然这里的方程看起来有点复杂，但是不必慌张.



图 7-14　鲁道夫·卡尔曼（1930—　）

定理 7.11　卡尔曼滤波器

$$\hat{X}(n)=A\hat{X}(n-1)+K_n[Y(n)-CA\hat{X}(n-1)] \tag{7.15}$$

$$K_n=S_nC'[CS_nC'+\Sigma_W]^{-1} \tag{7.16}$$

$$S_n=A\Sigma_{n-1}A'+\Sigma_V \tag{7.17}$$

$$\Sigma_n=(1-K_nC)S_n. \tag{7.18}$$

同时，

$$S_n = \mathrm{cov}(X(n) - A\hat{X}(n-1))\text{，}\quad \Sigma_n = \mathrm{cov}(X(n) - \hat{X}(n))\,. \qquad (7.19)$$

□

下面会给出很多应用这个结论的例子，首先来看一些说明.

- ❑ 方程（7.15）~（7.18）是递归的：在时刻 n 的估计值是关于时刻 n–1 估计值以及新观测值 $Y(n)$的简单线性函数.
- ❑ 矩阵 K_n 表示滤波器的增益, 可以在时刻 0 提前计算出来.
- ❑ 误差 $X(n)-\hat{X}(n)$的协方差 Σ_n 不依赖于观测值 $\{Y(0),\cdots,Y(n)\}$，因此可以在时刻 0 就提前计算出来. 估计值 $\hat{X}(n)$依赖于这些观测值, 但是误差并不依赖于这些观测值.
- ❑ 如果 $X(0)$与噪声随机变量都服从高斯分布,那么卡尔曼滤波器算出的是 MMSE.
- ❑ 最后可以看到,尽管这些方程有一些复杂,但在程序实现上只需要几行代码. 由于这一便利，卡尔曼滤波器在实际应用中广受推崇.

7.6.2　示例

本节讨论几个卡尔曼滤波器的例子.

1. 随机游走

卡尔曼滤波器的第一个实例是通过带噪声的观测值来追踪一个随机游走.

设

$$X(n+1) = X(n) + V(n) \qquad (7.20)$$

$$Y(n) = X(n) + W(n) \qquad (7.21)$$

$$\mathrm{var}(V(n)) = 0.04,\ \mathrm{var}(W(n)) = 0.09. \qquad (7.22)$$

也就是说，$X(n)$有相互正交的增量，并且在观测时带有互相正交的噪声．图 7-15 给出了这个滤波器的仿真．左图表明对状态的估计值很好地追踪了系统的状态，误差有限．中间的图给出了误差的方差．这一方差可以提前计算．右图给出了具有时变增益的滤波器（蓝色）以及具有稳态增益的滤波器（绿色）．考虑长期性能的时候，采用常数增益的滤波器与采用时变增益的滤波器表现都很好.

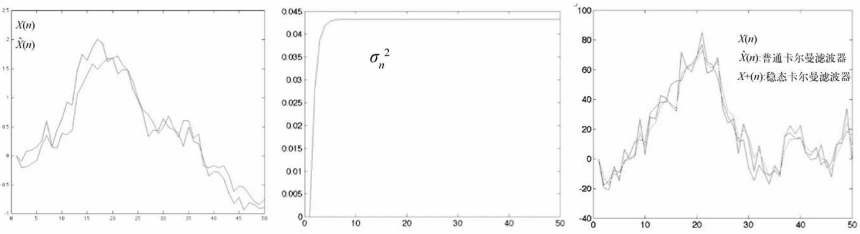


图 7-15　式（7.20）~（7.22）的卡尔曼滤波器（另见彩插）

2. 带未知漂移的随机游走

第二个例子研究具有未知漂移的随机游走．系统根据如下方程构成的模型进行演变：

$$X_1(n+1) = X_1(n) + X_2(n) + V(n) \tag{7.23}$$

$$X_2(n+1) = X_2(n) \tag{7.24}$$

$$Y(n) = X_1(n) + W(n) \tag{7.25}$$

$$\text{var}(V(n)) = 1, \text{var}(W(n)) = 0.25 . \tag{7.26}$$

在这个模型中，$X_2(n)$是恒定但未知的漂移，$X_1(n)$是随机游走的值．图 7-16 给出了滤波器的仿真．从图中可以看到，滤波器逐渐得到了漂移值，对于游走位置的估计也非常准确．

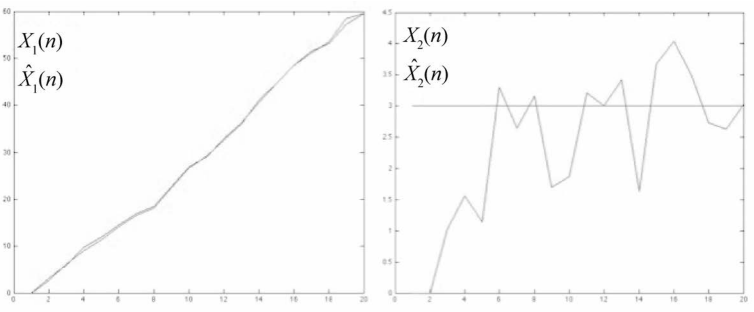


图 7-16　式（7.23）~（7.26）的卡尔曼滤波器（另见彩插）

3. 具有变化漂移的随机游走

第三个例子研究具有变化漂移的随机游走．系统的演变由如下方程决定：

$$X_1(n+1) = X_1(n) + X_2(n) + V_1(n) \tag{7.27}$$

$$X_2(n+1) = X_2(n) + V_2(n) \tag{7.28}$$

$$Y(n) = X_1(n) + W(n) \tag{7.29}$$

$$\text{var}(V_1(n)) = 1, \text{var}(V_2(n)) = 0.01 \tag{7.30}$$

$$\text{var}(W(n)) = 0.25 . \tag{7.31}$$

在这个模型中，$X_2(n)$ 是变化的漂移，$X_1(n)$ 是随机游走的值．图 7-17 给出了滤波器的仿真．此图表明滤波器尝试追踪漂移的值，并且对于游走的位置给出了准确的估计．

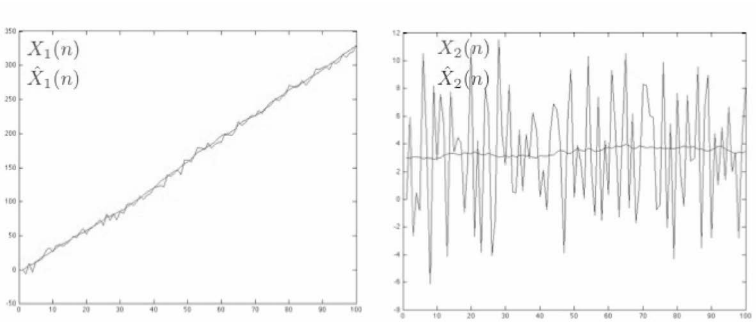


图 7-17　式（7.27）~（7.31）的卡尔曼滤波器（另见彩插）

4．自由下落的物体

第四个例子研究坠落物体的轨迹．自由下落物体的高度 $Z(n)$满足如下等式：

$$Z(n) = Z(0) + S(0)n - gn^2/2 + V(n), n \geqslant 0 .$$

其中 $S(0)$是物体沿垂直方向的初速度，g 是地球表面的万有引力常数．在这个表达式中，$V(n)$表示干扰运动的一些扰动．在每个时刻可以观察到 $\eta(n) = Z(n) + W(n)$，其中 $W(n)$是测量中的噪声．

由于 $-gn^2/2$ 是已知的，所以

$$X_1(n) = Z(n) + gn^2/2 \text{ 且 } Y(n) = \eta(n) + gn^2/2 .$$

通过上述变量的变换之后，系统的变化可以用如下的等式来描述：

$$X_1(n+1) = X_1(n) + X_2(n) + V(n) \tag{7.32}$$

$$X_2(n+1) = X_2(n) \tag{7.33}$$

$$Y(n) = X_1(n) + W(n) \tag{7.34}$$

$$\text{var}(V_1(n)) = 100, \text{var}(W(n)) = 1600 . \tag{7.35}$$

图 7-18 给出了滤波器的仿真．滤波器计算出 $\hat{X}_1(n)$ 的值，我们从中减去 $gt^2/2$ 来得到物体实际高度 $Z(n)$的估计值．

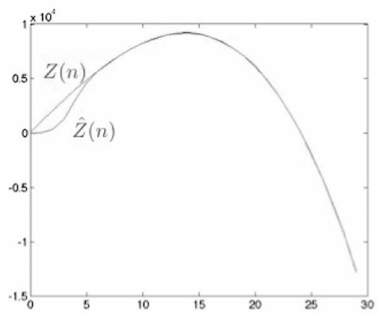


图 7-18　式（7.32）~（7.35）的卡尔曼滤波器（另件彩插）

7.7　小结

- ❑ LLSE，线性回归，MMSE
- ❑ 投影的特性
- ❑ 联合高斯分布的 MMSE 是线性的
- ❑ 卡尔曼滤波器

重要方程与公式

LLSE	$L[X\mid Y]=E(X)+\text{cov}(X,Y)\,\text{var}(Y)^{-1}(Y-E(Y))$	定理7.2
正交性	$X-L[X\mid Y]\perp a+bY$	式（7.6）
线性回归	收敛于 $L[X\mid Y]$	定理7.3
条件期望	$E[X\mid Y]=\cdots$	定义7.5
正交性	$X-E[X\mid Y]\perp g(Y)$	引理7.6
MMSE为条件期望	$MMSE[X\mid Y]=E[X\mid Y]$	定理7.4
条件期望的性质	线性性、光滑性等	定理7.7
联合高斯分布的条件期望	若X和Y服从联合高斯分布，则 $E[X\mid Y]=L[X\mid Y]=\cdots$	定理7.8
LLSE向量	$L[\boldsymbol{X}\mid\boldsymbol{Y}]=E(\boldsymbol{X})+\Sigma_{X,Y}\Sigma_Y^{-1}(\boldsymbol{Y}-E(\boldsymbol{Y}))$	定理7.10
卡尔曼滤波器	$\hat{X}(n)=A\hat{X}(n-1)+K_n[Y(n)-CA\hat{X}(n-1)]$	定理7.11

7.8　参考资料

LLSE、MMSE 和线性回归在 *Introduction to Probability* [Bertsekas and Tsitsiklis，2008]一书的第 4 章有所论述. 卡尔曼于 1960 年在 “A New Approach to Linear Filtering

and Prediction Problems”一文中首次提出并介绍了卡尔曼滤波器．Brown 和 Hwang 在 *Introduction to Random Signals and Applied Kalman Filtering* 一书中给出了有关于卡尔曼滤波器的入门知识以及很多例子．

7.9　练习

1．设 $X_n = Y_n + 2Y_n^2 + Z_n$，其中 Y_n 和 Z_n 是独立同分布于 $U[0, 1]$的随机变量．

(a) 计算 $L[X|Y]$以及 $E((X - L[X|Y])^2)$．

(b) 计算 $Q[X|Y]$以及 $E((X - Q[X|Y])^2)$，其中 $Q[X|Y]$是给定 Y 条件下 X 的二次方最小平方误差估计．

(c) 用 Matlab 实现一个能够计算 $Q[X|Y]$的随机梯度算法并计算 $Q[X|Y]$．

2．比较用离线方法和在线方法计算 $L[X|Y]$的不同．参数和变量的设定与题 1 一致．

(a) 产生 $N = 1000$ 个样本点，计算给定 Y 条件下 X 的线性回归．将答案记作 $X = aY + b$．

(b) 给定同样的样本点，用随机化的梯度算法来递归计算线性拟合的值．将所得答案记为 $X = cY + d$．

(c) 通过计算 $E((X - aY - b)^2)$ 和 $E((X - cY - d)^2)$ 来比较两个结果的优劣．

3．随机变量 X、Y、Z 服从联合高斯分布，

$$(X, Y, Z)^T \sim N((0,0,0)^T, \begin{bmatrix} 2 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}).$$

(a) 计算 $E[X | Y, Z]$．

(b) 计算误差的方差．

4．根据分布 $f_{X|\theta}(x) = \frac{1}{2}\mathrm{e}^{-|x-\theta|}$ $(\theta \in \mathbb{R})$采样三个独立同分布的样本点 X_1、X_2 和 X_3．其中 θ 是需要估值的参数．计算 $MLE[\theta | X_1, X_2, X_3]$．

5．(a) 给定三个独立同分布于 $N(0, 1)$的随机变量 X、Y、Z，求下式在最小均方误差下的估计值：$E[X+3Y|2Y+5Z]$．

(b) 求上式估计值的均方误差．

6．给定两个相互独立的 $N(0, 1)$随机变量 X 和 Y，求下式的线性最小平方估计：

$$L[X | X^2 + Y].$$

提示：服从 $N(0, 1)$分布的随机变量 X 的特征函数为 $E(\mathrm{e}^{isX}) = \mathrm{e}^{-\frac{1}{2}s^2}$．

7．考虑具有 n 个传感器的传感网络，对于信号 X 有一系列观测值 $\boldsymbol{Y}^n = (Y_1, \cdots, Y_n)$，其中

$$Y_i = aX + Z_i (i = 1, \cdots, n).$$

在上式中，$X =_D N(0,1)$ 且 $Z_i =_D N(0,\sigma^2)$（$i = 1$，$\cdots$，n），并且这些随机变量是相互独立的.

(a) 计算给定 $\boldsymbol{Y}^n$ 时 X 的 MMSE 估计.

(b) 计算这个估计值的均方误差 ${\sigma_n}^2$.

(c) 假设每次观测所付出的代价为 C. 若希望

$$nC + {\sigma_n}^2$$

最小，计算 n 的最优值.

(d) 假设在每一步都可以决定再进行一次测量或者停止. 我们要使得 $vC + {\sigma_v}^2$ 的期望值最小，其中的随机变量 v 是测量的次数. 是否有一种决策方法能够比(c)中得出的定值 n 更好？请解释你的结论.

8. 为了用卡尔曼滤波器来检测微博中某个词语的使用频率变化，设第 n 天这个词语在微博中出现了 $Y(n)$ 次，建模如下：

C.7

$$X(n+1) = X(n)$$
$$Y(n) = X(n) + W(n),$$

其中 $W(n)$的均值为 0，彼此不相关. 这个模型的意思是观测到事件发生了很多次，发生次数的均值 $X(n)$是未知的常数. 因此，如果均值在变化，就能够通过观察 $Y(n)$ 与 $\hat{Y}(n)$ 之间的误差变大来检测到. 提出一个算法来检测这种变化，并在 Matlab 中进行仿真.

9. 随机变量 X 服从指数分布，均值为 1. 随机变量 Y 呈指数分布，参数为 X.

(a) 计算 $E[Y|X]$.

(b) 计算 $E[X|Y]$.

10. X、Y、Z 为独立同分布的 $\mathcal{N}(0, 1)$变量

(a) 计算 $L[X^2 + Y^2 \mid X + Y]$.

(b) 计算 $E[X + 2Y \mid X + 3Y + 4Z]$.

(c) 计算 $E[(X + Y)^2 \mid X - Y]$.

11. 设 $(V_n, n \geqslant 0)$ 独立同分布于 $N(0,\sigma^2)$ ，且与 $X_0 = N(0,u^2)$ 独立. 定义

$$X_{n+1} = aX_n + V_n, n \geqslant 0.$$

(a) 求 X_n ($n \geqslant 1$)的分布.

(b) 计算 $E[X_{n+m} \mid X_n]$ ($m \leqslant n < n + m$).

(c) 计算使得所有 X_n 具有相同分布的 u 值.

12. 设 $\theta =_D U[0,1]$. 给定 θ ，随机变量 X 均匀分布在 $[0,\theta]$ 上. 计算 $E[\theta \mid X]$.

13. 设 $(X,Y)^T \sim N([0;0],[3,1;1,1])$. 求 $E[X^2 \mid Y]$.

14. 设 $(X,Y,Z)^T \sim N([0;0;0],[5,3,1;3,9,3;1,3,1])$. 求 $E[X\mid Y,Z]$.

15. 对于任意的随机变量 X 和 Y，证明如下的性质：

$$\operatorname{var}(Y)=E(\operatorname{var}(Y\mid X))+\operatorname{var}(E(Y\mid X)).$$

16. 设两个随机变量 X 和 Y 的联合密度为

$$f_{X,Y}(x,y)=\frac{1}{4}(2x+y)1\{0\leqslant x\leqslant 1\}1\{0\leqslant y\leqslant 2\}.$$

首先证明这是一个合理的联合概率密度函数．然后，假设从这个联合分布中观测到了 Y 的值，求 $MMSE[X|Y]$.

17. 给定四个相互独立、服从 $N(0, 1)$分布的随机变量 X、Y、Z 和 V，求下式的最小方均误差估计值：

$$E[X+2Y+3Z\mid Y+5Z+4V].$$

然后，求这个估计值的均方误差.

18. 设 X 和 Y 是满足 $E[X\mid Y]=L[X\mid Y]$ 的随机变量．那么一定有：

❑ X 和 Y 服从联合高斯分布；

❑ X 可以写作 $X=aY+Z$，其中 Z 是独立于 Y 的随机变量；

❑ $E((X-L[X\mid Y])Y^k)=0$ 对于所有的 $k\geqslant 0$ 成立；

❑ $E((X-L[X\mid Y])\sin(3Y+5))=0$.

19. 在具有高斯噪声的线性系统中，系统状态为 X_n，观测值为 Y_n，卡尔曼滤波器能够计算（请选择正确答案，$Y^n=(Y(1),\cdots,Y(n))$）：

❑ $MLE[Y_n\mid X^n]$

❑ $MLE[X_n\mid Y^n]$

❑ $MAP[Y_n\mid X^n]$

❑ $MAP[X_n\mid Y^n]$

❑ $E[X_n\mid Y^n]$

❑ $E[Y_n\mid X^n]$

❑ $E[X_n\mid Y_n]$

❑ $E[Y_n\mid X_n]$

20. 设（X, $\boldsymbol{Y}$）服从 $N(\mu,\Sigma)$，其中 $\mu'=[2,1,3,4,5]$，$\boldsymbol{Y}'=[Y_1,Y_2,Y_3,Y_4]$，并且

$$\Sigma=\begin{bmatrix}3&4&6&12&8\\4&6&9&18&12\\6&9&14&28&18\\12&18&28&56&36\\8&12&18&36&24\end{bmatrix}.$$

计算 $E[X|\boldsymbol{Y}]$.

21. 设 $\boldsymbol{X} = A\boldsymbol{V}$，$\boldsymbol{Y} = C\boldsymbol{V}$，其中 $\boldsymbol{V} = N(\boldsymbol{0}, \boldsymbol{I})$，计算 $E[\boldsymbol{X}|\boldsymbol{Y}]$.

22. 给定 $\theta \in \{0,1\}$ ，$\boldsymbol{X} = N(\boldsymbol{0}, \Sigma_\theta)$ ，其中

$$\Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{且} \Sigma_1 = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

$\rho > 0$ 为给定值.

计算 $MLE[\theta \mid \boldsymbol{X}]$.

23. 给定两个独立分布为 $N(0, 1)$的随机变量 X 和 Y,计算下式的线性最小平方估计：

$$L[X \mid X^3 + Y].$$

提示：服从 $N(0, 1)$分布的随机变量 X 的特征函数为

$$E(\mathrm{e}^{isX}) = \mathrm{e}^{-\frac{1}{2}s^2}.$$

24. 设 X、Y、Z 独立同分布于 $\mathcal{N}(0, 1)$. 计算

$$E[X \mid X+Y, X+Z, Y-Z].$$

提示：请说明观测值 $Y–Z$ 是多余的.

25. 设 X、Y_1、Y_2 和 Y_3 的均值都是 0，协方差矩阵为

$$\Sigma = \begin{bmatrix} 10 & 6 & 5 & 16 \\ 6 & 9 & 6 & 21 \\ 5 & 6 & 6 & 18 \\ 16 & 21 & 18 & 57 \end{bmatrix}.$$

计算 $L[X|Y_1, Y_2, Y_3]$.

提示：你会发现 $\Sigma_{\boldsymbol{Y}}$ 是一个奇异阵. 这意味着至少一个观测值（Y_1、Y_2 或 Y_3）是多余的，因为它是其他两个观测值的线性组合. 这意味着 $L[X|Y_1, Y_2, Y_3] = L[X|Y_1, Y_2]$.

第8章

追踪定位—B

主题：卡尔曼滤波器的推导及性质；扩展卡尔曼滤波器

背景知识：第 7 章

在第 7 章中，我们解释了如何通过观察对随机变量进行估计，并且介绍了卡尔曼滤波器及一些例子. 本章将对卡尔曼滤波器进行推导并阐述它的一些性质，还将进一步讨论扩展卡尔曼滤波器.

8.1 LLSE的更新

在很多情况下，我们会不断地进行观察并希望依据观察得到的数据对随机变量的估计进行更新. 此时，我们希望能利用一些已有的结果，而不是每次都从头算起. 也就是说，我们希望在知道 $L[\boldsymbol{X}\mid\boldsymbol{Y}]$ 和 $\boldsymbol{Z}$ 的情况下，能够高效地计算出 $L[\boldsymbol{X}\mid\boldsymbol{Y},\boldsymbol{Z}]$.

下面的定理给出了这个方法的主要思想.

定理 8.1　LLSE 的更新——加性正交化观测

设 $\boldsymbol{X}$、$\boldsymbol{Y}$、$\boldsymbol{Z}$ 为均值为 0 的随机变量，并且 $\boldsymbol{Y}$ 与 $\boldsymbol{Z}$ 互相正交. 那么，

$$L[\boldsymbol{X}\mid\boldsymbol{Y},\boldsymbol{Z}]=L[\boldsymbol{X}\mid\boldsymbol{Y}]+L[\boldsymbol{X}\mid\boldsymbol{Z}]. \tag{8.1}$$

证明　图 8-1 形象地解释了这个定理为何成立，下面在数学上给予证明. □

我们需要证明误差

$$\boldsymbol{X}-(L[\boldsymbol{X}\mid\boldsymbol{Y}]+L[\boldsymbol{X}\mid\boldsymbol{Z}])$$

与 $\boldsymbol{Y}$、$\boldsymbol{Z}$ 都互相正交. 首先注意到这个误差可以写成以下的形式：

$$(\boldsymbol{X}-L[\boldsymbol{X}\mid\boldsymbol{Y}])-L[\boldsymbol{X}\mid\boldsymbol{Z}].$$

根据 $L[\boldsymbol{X}\mid\boldsymbol{Y}]$ 的投影性质，括号中的式子与 $\boldsymbol{Y}$ 正交. 第二项 $L[\boldsymbol{X}\mid\boldsymbol{Z}]$ 是 $\boldsymbol{Z}$ 中元素的线性组合. 因为 $\boldsymbol{Z}$ 与 $\boldsymbol{Y}$ 正交，所以第二项也垂直于 $\boldsymbol{Y}$. 由此整个误差与 $\boldsymbol{Y}$ 正交. 同理可

证，误差与 $\boldsymbol{Z}$ 正交. ♣

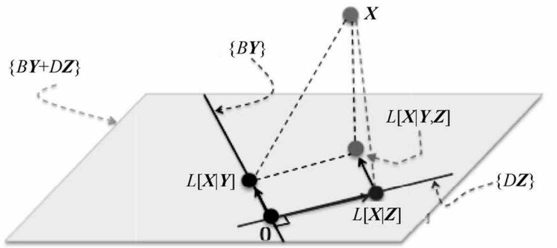


图 8-1 在加性正交化观测中，LLSE 很容易更新

$$\boldsymbol{X}-(L[\boldsymbol{X}\mid\boldsymbol{Y}]+L[\boldsymbol{X}\mid\boldsymbol{Z}])$$

由定理 8.1 很容易推出如下结论.

定理 8.2 LLSE 更新——一般情况下的加性观测

设 $\boldsymbol{X}$、$\boldsymbol{Y}$、$\boldsymbol{Z}$ 均值为 0，则

$$L[\boldsymbol{X}\mid\boldsymbol{Y},\boldsymbol{Z}]=L[\boldsymbol{X}\mid\boldsymbol{Y}]+L[\boldsymbol{X}\mid\boldsymbol{Z}-L[\boldsymbol{Z}\mid\boldsymbol{Y}]]. \tag{8.2}$$

□

证明 在新的观测量当中，与 $\boldsymbol{Y}$ 垂直的部分是 $\tilde{\boldsymbol{Z}}=\boldsymbol{Z}-L[\boldsymbol{Z}\mid\boldsymbol{Y}]$，它正好也包含了需要更新的信息.

下面来说明为何该结论成立. 首先，任何 $\boldsymbol{Y}$ 和 $\boldsymbol{Z}$ 的线性组合都可以表示为 $\boldsymbol{Y}$ 与 $\tilde{\boldsymbol{Z}}$ 的线性组合. 例如，如果 $L[\boldsymbol{Z}\mid\boldsymbol{Y}]=C\boldsymbol{Y}$，则

$$A\boldsymbol{Y}+B\boldsymbol{Z}=A\boldsymbol{Y}+B(\boldsymbol{Z}-C\boldsymbol{Y})+BC\boldsymbol{Y}=(A+BC)\boldsymbol{Y}+B\tilde{\boldsymbol{Z}}.$$

因此，由关于 $\boldsymbol{Y}$ 和 $\boldsymbol{Z}$ 的线性方程组成的集合与关于 $\boldsymbol{Y}$ 和 $\tilde{\boldsymbol{Z}}$ 的相同. 所以

$$L[\boldsymbol{X}\mid\boldsymbol{Y},\boldsymbol{Z}]=L[\boldsymbol{X}\mid\boldsymbol{Y},\tilde{\boldsymbol{Z}}].$$

由定理 8.1 以及 $\boldsymbol{Y}$ 和 $\tilde{\boldsymbol{Z}}$ 相互正交可证式（8.2）.

♣

8.2 卡尔曼滤波器的推导

下面来推导定理 7.11 中用到的卡尔曼滤波器方程. 为便于本节说明，将其重写如下：

$$\hat{X}(n)=A\hat{X}(n-1)+K_n[Y(n)-CA\hat{X}(n-1)] \tag{8.3}$$

$$K_n=S_nC'[CS_nC'+\Sigma_W]^{-1} \tag{8.4}$$

$$S_n=A\Sigma_{n-1}A'+\Sigma_V \tag{8.5}$$

$$\Sigma_n = (1 - K_n C) S_n \tag{8.6}$$

以及

$$S_n = \text{cov}(X(n) - A\hat{X}(n-1)), \quad \Sigma_n = \text{cov}(X(n) - \hat{X}(n)). \tag{8.7}$$

在计算中，我们会反复用到下面这一性质：

$$\text{cov}(BV, DW) = B\,\text{cov}(V, W) D',$$

以及，如果 V 和 W 互相正交，则

$$\text{cov}(V + W) = \text{cov}(V) + \text{cov}(W).$$

下面的推导有点繁琐，但是思路简单明确.

设

$$Y^n = (Y(0), \cdots, Y(n)),$$

由式（7.13）可得

$$L[X(n) \mid Y^{n-1}] = L[AX(n-1) + V(n-1) \mid Y^{n-1}] = A\hat{X}(n-1).$$

由式（7.14）可以导出

$$L[Y(n) \mid Y^{n-1}] = L[CX(n) + W(n) \mid Y^{n-1}] = CL[X(n) \mid Y^{n-1}] = CA\hat{X}(n-1).$$

因此，

$$Y(n) - L[Y(n) \mid Y^{n-1}] = Y(n) - CA\hat{X}(n-1).$$

由定理 8.2 可得

$$\begin{aligned}
\hat{X}(n) &= L[X(n) \mid Y^n] = L[X(n) \mid Y^{n-1}] + L[X(n) \mid Y(n) - L[Y(n) \mid Y^{n-1}]] \\
&= A\hat{X}(n-1) + K_n[Y(n) - CA\hat{X}(n-1)].
\end{aligned}$$

从推导中可以看到，式（8.3）是定理 8.2 中对 LLSE 数据更新公式的直接推论.

对于增益 K_n 的推导则有点复杂．设

$$\tilde{Y}(n) = Y(n) - L[Y(n) \mid Y^{n-1}] = Y(n) - CA\hat{X}(n-1),$$

现在看到

$$K_n = \text{cov}(X(n), \tilde{Y}(n))\,\text{cov}(\tilde{Y}(n))^{-1},$$

其中，由于 $\tilde{Y}(n)$ 与 Y^{n-1} 相互垂直，

$$\text{cov}(X(n), \tilde{Y}(n)) = \text{cov}(X(n) - L[X(n) \mid Y^{n-1}], \tilde{Y}(n)).$$

现在，根据式（8.7）可得

$$\begin{aligned}
&\text{cov}(X(n) - L[X(n) \mid Y^{n-1}], \tilde{Y}(n)) \\
&\quad = \text{cov}(X(n) - A\hat{X}(n-1), Y(n) - CA\hat{X}(n-1)) \\
&\quad = \text{cov}(X(n) - A\hat{X}(n-1), CX(n) + W(n) - CA\hat{X}(n-1)) \\
&\quad = S_n C'.
\end{aligned}$$

下面计算 $\text{cov}(\tilde{Y}(n))$，

$$\text{cov}(\tilde{Y}(n)) = \text{cov}(CX(n) + W(n) - CL[X(n) \mid Y^{n-1}]) = CS_n C' + \Sigma_W.$$

由此得到，

$$K_n = S_n C'[CS_n C' + \Sigma_W]^{-1}.$$

为了证明式（8.5），注意到

$$\begin{aligned} S_n &= \text{cov}(X(n) - L[X(n) \mid Y^{n-1}]) \\ &= \text{cov}(AX(n-1) + V(n-1) - A\hat{X}(n-1)) \\ &= A\Sigma_{n-1}A' + \Sigma_V. \end{aligned}$$

最后推导式（8.6）. 观察到

$$\begin{aligned} X(n) - L[X(n) \mid Y^n] &= X(n) - A\hat{X}(n-1) - K_n[Y(n) - CA\hat{X}(n-1)] \\ &= X(n) - A\hat{X}(n-1) - K_n[CX(n) + W(n) - CA\hat{X}(n-1)] \\ &= [I - K_n C][X(n) - A\hat{X}(n-1)] - K_n W(n), \end{aligned}$$

所以，由 $\Sigma_n = \text{cov}(X(n) - \hat{X}(n))$ 可得

$$\begin{aligned} \Sigma_n &= [I - K_n C]S_n[I - K_n C]' + K_n \Sigma_W K_n' \\ &= S_n - 2K_n CS_n + K_n[CS_n C' + \Sigma_W]K_n' \\ &= S_n - 2K_n CS_n + K_n[CS_n C' + \Sigma_W][CS_n C' + \Sigma_W]^{-1}CS_n \text{（由式（8.4）得到）} \\ &= S_n - K_n CS_n. \end{aligned}$$

结论得证.

8.3 卡尔曼滤波器的特性

本节的目标是解释并证明定理 8.3. 我们首先给出定理的内容，然后给出可观测和可达这两个概念的定义.

定理 8.3 卡尔曼滤波器的特性

(a) 若(A,C)是可观测的，则 Σ_n 有界. 若同时满足 $\Sigma_0 = \mathbf{0}$，则

$$\Sigma_n \to \Sigma \text{ 且 } K_n \to K, \tag{8.8}$$

其中 Σ 是有限维的矩阵.

(b) 如果 $(A, \Sigma_V^{1/2})$ 在满足上述条件的同时还满足可达到性，则可采用 $K_n = K$ 使得误差的协方差收敛于 Σ. □

我们在后续小节中将陆续解释这些性质. 现在先给出一些说明.

- 在有些系统中，误差会无限地增大．例如，如果没有任何观测值（如 $C = 0$）并且系统不稳定（如 $X(n) = 2X(n-1)+V(n)$），则 Σ_n 会趋于无穷大．不过，(a)中"在观测量足够丰富的情况下"的条件已经保证了无穷大的情况不会出现．这一点会在后面详细讨论．
- 定理的(b)部分指出在一些情况下，即使在滤波器中采用一个固定增益 K，误差也不会在渐进过程中变大．这一点非常有用，因为不需要在每一步都计算一个新的增益．

8.3.1 可观测性

如何知道观测值能否在追踪定位时将误差的协方差控制在一个范围内？在给出结论以前，我们先对观测值的好坏给出准确的定义．

定义 8.4 可观测性

如果矩阵

$$\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{d-1} \end{bmatrix}$$

的零空间为 $\{\mathbf{0}\}$，则 (A, C) 是**可观测**的．这里，d 是 $X(n)$ 的维数．一个矩阵 M 的零空间为 $\{\mathbf{0}\}$，意味着唯一能使 $Mv = \mathbf{0}$ 的向量 v 就是零向量． ◊

下面是一个重要的结论．

引理 8.5 可观测性等价于有界的误差方差

(a) 如果系统是可观测的，则 Σ_n 有界.

(b) 如果同时满足 $\Sigma_0 = 0$，则 Σ_n 收敛于某个有限的值 Σ.

证明

(a) 可观测性意味着在没有噪声的情况下，观测值 $(Y(0),\cdots,Y(d-1))$ 只对应一个 $X(0)$．这时，

$$X(n) = AX(n-1), Y(n) = CX(n).$$

所以，

$$X(1) = AX(0), X(2) = A^2X(0), \cdots, X(d-1) = A^{d-1}X(0).$$

因此，

$$Y(0)=CX(0), Y(1)=CAX(0), \cdots, Y(d-1)=CA^{d-1}X(0).$$

由此可得，

$$\begin{bmatrix} Y(0) \\ Y(1) \\ \vdots \\ Y(d-1) \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{d-1} \end{bmatrix} X(0).$$

如果有 2 种初始状态，分别记为 $X(0)$和$\mathring{X}(0)$，都具有相同的输出 $Y(0),\cdots,Y(d-1)$. 则

$$\begin{bmatrix} Y(0) \\ Y(1) \\ \vdots \\ Y(d-1) \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{d-1} \end{bmatrix} X(0) = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{d-1} \end{bmatrix} \mathring{X}(0),$$

所以

$$\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{d-1} \end{bmatrix} (X(0)-\mathring{X}(0)) = \mathbf{0}.$$

根据可观测性的定义，得出 $X(0)-\mathring{X}(0)=\mathbf{0}$.

因此，如果我们根据 d+1 个输出能够唯一地确定出初始状态 $X(0)$，那么(A,C)是可观测的. 同时因为没有噪声，根据这个初值就可以准确地知道 $X(1), X(2),\cdots$ 的值. 因此，当(A,C)可观测时，可以根据输出的值准确无误地得到 $X(n)$的值.

但是事实上，我们的系统总是有噪声的. 如果(A,C)可观测，就可以根据 $Y(0),\cdots,Y(d-1)$ 估计 $X(0)$的值. 不过这里得到的估值并不完全准确，而是带有一个关于噪声的线性函数，即关于$\{V(0),\cdots,V(d-1),W(0),\cdots,W(d)\}$ 的线性函数. 于是，也可以根据 $Y(0),\cdots,Y(d-1)$ 估计 $X(d-1)$ 的估值，这个值也带有一个关于 $\{V(0),\cdots,V(d-1),W(0),\cdots,W(d)\}$ 的线性函数. 同样道理，可以从 $Y(n-d),\cdots,Y(n)$ 得到一个关于 $X(n)$的估计，它带有关于$\{V(n),\cdots, V(n+d-1),W(n-d),\cdots,W(n)\}$ 的线性函数.

上述论证也表明 $X(n)$与 $\hat{X}(n)$之间的误差只由有限的误差值决定，所以 Σ_n 是有界的.

(b) 如果 $\Sigma_0=0$，即知道 $X(0)$，则能证明 $\Sigma_n-\Sigma_{n-1}$ 是非负矩阵. 因此 Σ_n 总是递增的. 有界且递增说明 Σ_n 最终收敛. 同理可知，K_n 也是收敛的. ♣

8.3.2 可达性

假设 $\Sigma_V = QQ'$. 如果矩阵

$$[Q, AQ, \cdots, A^{d-1}Q]$$

满秩，则称(A,Q)是可达的．为了让大家更深刻地认识到这个结论的意义，我们将已有的等式写作

$$X(n) = AX(n-1) + Q\eta_n\ ,$$

其中$\mathrm{cov}(\eta_n) = \boldsymbol{I}$．也就是说，$\eta$中的元素彼此正交．可以设$\eta$中的元素服从 $N(0,1)$分布而且相互独立. 如果(A,Q)是可达的,则对于任意的$\boldsymbol{x} \in \Re^d$,都存在一个序列$\eta_0,\cdots,\eta_d$ 使得如果 $X(0) = 0$，则 $X(d) = \boldsymbol{x}$．事实上，

$$X(d) = \sum_{k=0}^{d} A^k Q\eta_{d-k} = [Q, AQ,\cdots, A^{d-1}Q]\begin{bmatrix}\eta_d \\ \eta_{d-1} \\ \vdots \\ \eta_0\end{bmatrix}.$$

由于矩阵是满秩的，它的列空间为$\Re^d$．也就是说$\Re^d$中的任意向量都可以用这个矩阵中列向量的线性组合来表示.

定理(b)部分的证明太过复杂，本书从略.

8.4 扩展卡尔曼滤波器

卡尔曼滤波器常被用于非线性系统．主要思路是，如果系统很快进入了一个几乎线性的状态，那么可以在其局部使用卡尔曼滤波器并根据状态估值的变化调整 A 和 C 矩阵.

这里的系统模型如下：

$$X(n+1) = f(X(n)) + V(n)$$

$$Y(n+1) = g(X(n+1)) + W(n+1).$$

扩展卡尔曼滤波器为

$$\begin{aligned}
&\hat{X}(n+1) = f(\hat{X}(n)) + K_n[Y(n+1) - g(f(\hat{X}(n)))] \\
&K_n = S_n C'_n [C_n S_n C'_n + \Sigma_W]^{-1} \\
&S_n = A_n \Sigma_n A'_n + \Sigma_V \\
&\Sigma_{n+1} = [I - K_n C_n] S_n,
\end{aligned}$$

其中

$$[A_n]_{ij} = \frac{\partial}{\partial x_j} f_i(\hat{X}(n))\ ,\quad [C_n]_{ij} = \frac{\partial}{\partial x_j} g_i(\hat{X}(n)).$$

这里的思想就是在系统状态的估计值附近用线性来近似，然后应用普通的卡尔曼滤波器.

我们已经进入了启发式方法的领域．在这种情况下，难以对于滤波器的性质作出

预测．实验结果说明，当系统非线性很弱的时候，这种方法的效果良好．但是在其他情况下可能会出现极大的误差．

示例

1．追踪定位车辆

这个例子来自于“Eric Feron, Notes for AE6531, Georgia Tech”教案．例子研究的是一个追踪定位车辆的问题．假设有一辆小车在平面内移动，我们通过测量小车到 9 个平面上的点 $p_i \in \Re^2$ 的距离对其进行定位（测量会有误差）．设 $p(n) \in \Re^2$ 是小车在 n 时刻的位置，记 $u(n) \in \Re^2$ 为它的速度．

假设速度的变化主要服从于一个已知的规律，但存在随机的扰动．具体来说，我们假设

$$p(n+1) = p(n) + 0.1u(n), \tag{8.9}$$

$$u(n+1) = \begin{bmatrix} 0.85 & 0.15 \\ -0.1 & 0.85 \end{bmatrix} u(n) + w(n), \tag{8.10}$$

其中 $w(n)$为独立同分布的 $N(0,\boldsymbol{I})$变量．测量值为

$$y_i(n) = \|p(n) - p_i\| + v_i(n), i = 1, 2, \cdots, 9.$$

其中 $v_i(n)$ 为独立同分布的 $N(0, 0.3^2)$ 变量．

图 8-2 给出了扩展卡尔曼滤波器的结果示意图，其中 $X(n) = (p(n), u(n))$ 的初始值为 $\hat{X}(0) = 0$，$\Sigma_0 = \boldsymbol{I}$．

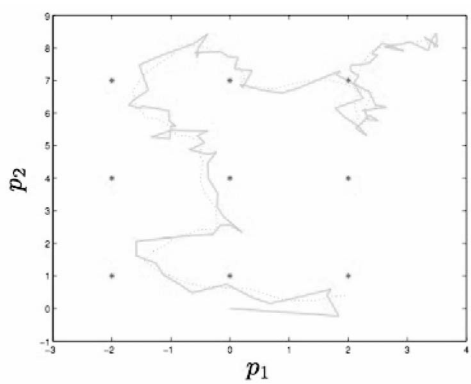


图 8-2　扩展卡尔曼滤波器用于式（8.9）～（8.10）描述的系统（另见彩插）

2. 观测化学反应

这个例子讲述了如何通过气压的测量值来估计化学反应进行的状态．这个例子来自 James B. Rawlings 和 Fernando V. Lima、U. Wisconsin、Madison．反应中有 A、B、C 三种物质，它们之间化学反应的模型如图 8-3 所示，其中 k_i 是动力学常数．

$$\boxed{\begin{array}{l} \mathrm{A} \underset{k-1}{\overset{k_1}{\rightleftharpoons}} \mathrm{B}+\mathrm{C} \\ 2\mathrm{B} \underset{k-2}{\overset{k_2}{\rightleftharpoons}} \mathrm{C} \end{array}}$$

图 8-3　化学反应

设 C_A、C_B、C_C 分别表示三种物质的浓度．系统模型为：

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} C_A \\ C_B \\ C_C \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 1 & -2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} k_1 C_A - k_{-1} C_B C_C \\ k_2 C_B^2 - k_{-2} C_C \end{bmatrix}$$

以及

$$y = RT(C_A + C_B + C_C).$$

如图 8-4 上面小图所示，这个滤波器并不能正确追踪出浓度数值．事实上，有一些浓度的估计值甚至是负的！

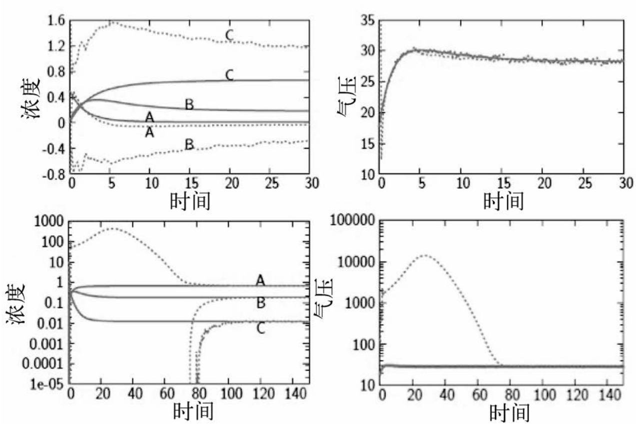


图 8-4　上方的两幅小图表明扩展卡尔曼滤波器并没有正确地追踪出浓度．下方的两幅小图表明在修改了方程以后，算法收敛到正确的答案（另见彩插）

下方的小图则说明，在修改了方程并将负值替换为 0 以后，得到的浓度值变得收敛了．

举这个例子是为了说明，扩展卡尔曼滤波器并不总能保证收敛，但有时简单的修改就能使滤波器收敛．

8.5 小结

- ❑ LLSE 估值更新
- ❑ 卡尔曼滤波器的推导
- ❑ 可观测性与可达性
- ❑ 扩展卡尔曼滤波器

重要方程与公式

LLSE估值更新（在均值为0的情况下）	$\Rightarrow L[\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{Z}] = L[\boldsymbol{X} \mid \boldsymbol{Y}] + L[\boldsymbol{X} \mid \boldsymbol{Z} - L[\boldsymbol{Y} \mid \boldsymbol{Z}]]$	定理8.2
可观测性	$\Rightarrow$ 误差的方差有界	引理8.5
可观测性与可达性	$\Rightarrow$ 渐进的滤波器就足够好	定理7.11
扩展卡尔曼滤波器	方程的线性化	8.4节

8.6 参考资料

Goodwin 和 Sin 的 *Adaptive Filtering Prediction and Control* 一书综述了滤波器和控制理论的相关应用．Kumar 和 Varaiya 所著的教科书 *Stochastic Systems: Estimation, Identification and Adaptive Control* 全面地展示了控制理论、滤波器和自适应控制．该教材难度适中，可以通过网络资源获取．

第9章

语音识别—A

应用：估计，假设检验，语音识别
主题：隐马尔可夫链，维特比解码，期望最大化算法

背景知识：第 1 章，第 5 章，第 7 章



图 9-1　你能听见我吗

语音识别可以被规范为这样一个问题：猜测一系列声音所对应的单词. 人类大脑对此十分擅长，即便同一个单词可能因为口音或声音特点不同而对应不同的发音，大脑也总能辨认. 不仅如此，由于背景总是带有噪声，所以听者听到的总是衰损版的语音.

如今计算机在语音识别方面做得越来越好. 语音控制系统如今在智能手机（Siri）、汽车（GPS、音乐和气温控制）、电话呼叫中心和录音系统中都很常见. 本章会介绍语音识别算法和其他相关应用的主要思想.

我们首先建立一个系统模型，包含一个待检测的随机序列（比如单词），及其与观测到的信号（比如语音）之间的关联. 主要的模型叫作*隐马尔可夫链*. 主要思路是，连续的音节组成一个马尔可夫链，且每一个单词随机映射到一些音节上. 这一模型也在信系统中被用于解码符号串.

下面来探讨一些有关学习的基本概念.

9.1　学习：概念和示例

在人工智能中，“学习”指的是从关联的事物中寻找联系的过程.

举一个简单的例子，考虑第 5 章问题 5 中的二元对称信道．输入 X_n 是服从 $B(p)$ 的独立同分布随机变量．对给定的输入值 X_n（$n\geqslant 0$），输出值 Y_n 等于输入 X_n 的概率为 $1-\varepsilon$．在这个例子中，输入和输出间有一个由 ε 决定的概率性关联．学习在这里意味着估计 ε 的值．

学习有两种基本情况．在监督学习中，观测输入$\{X_n, n=0, \cdots, N\}$和输出$\{Y_n, n=0, \cdots, N\}$．可以将这一步骤看作系统的训练阶段．因此，先用一系列已知的输入值来观测信道．一旦“学习”了信道，也就是估计出了 ε 的值，就可以着手设计最佳的接收机并将其用于未知的输入．在无监督学习中，我们只观测输出．这种学习方式的优点在于，学习过程并不影响系统运作．因此，训练阶段并不会浪费时间．而且，如果 ε 的值变化不快，系统还可以自动调整估计值，无需在新的训练阶段重新学习它．

不难看到，在选择监督学习和无监督学习时需要有所权衡．对系统进行训练需要花费时间，但是学习速度通常比无监督学习更快．在实际中，我们总是根据系统的特点，如系统参数改变的速率等，来选择最适合的方法．

9.2 隐马尔可夫链

隐马尔可夫链是一个带有状态观测模型的马尔可夫链．一个马尔可夫链$\{X(n), n\geqslant 0\}$建立在状态空间 $\mathcal{X}$ 上，并有自己的转移矩阵 P 和初始分布 π_0．而状态观测模型指明，当马尔可夫链处于状态 x 时，我们观测到值 y 的概率为 $Q(x, y)$，$y\in\mathcal{Y}$．下面给出正式的定义．

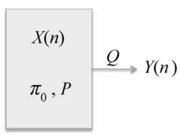


图 9-2　隐马尔可夫链

定义 9.1　隐马尔可夫链

隐马尔可夫链是一个满足 $X(n)\in\mathcal{X}=\{1, \cdots, N\}$，且 $Y(n)\in\mathcal{Y}=\{1, \cdots, M\}$ 的随机序列$\{(X(n), Y(n)), n\geqslant 0\}$，且对于所有 $n\geqslant 0$ 有 $x_n\in\mathcal{X}$，$y_n\in\mathcal{Y}$，

$$P(X(0)=x_0, Y(0)=y_0, \cdots, X(n)=x_n, Y(n)=y_n)$$
$$=\pi_0(x_0)Q(x_0, y_0)P(x_0, x_1)Q(x_1, y_1)\times\cdots\times P(x_{n-1}, x_n)Q(x_n, y_n) \tag{9.1}$$

都成立．　◊

在语音识别应用中，X_n 是语音的一部分，也就是句子的片段；而 Y_n 是声音．语言结构会决定建模所用的马尔可夫链中 X_n 之间的关系．X_n 和 Y_n 之间的关系则取决于说话者．

现在，识别问题可以准确表述如下．假设观察到 $\boldsymbol{Y}^n := (Y_0, \cdots, Y_n) = \boldsymbol{y}^n := (y_0, \cdots, y_n)$．那么哪个 $\boldsymbol{X}^n := (X_0, \cdots, X_n)$序列最可能是说话者表达的内容？如果采用第 5 章的术语，也就是要计算

$$MAP[\boldsymbol{X}^n|\boldsymbol{Y}^n = \boldsymbol{y}^n].$$

因此，我们想要找到序列 $\boldsymbol{x}^n \in \mathcal{X}^{n+1}$，使得

$$P[\boldsymbol{X}^n = \boldsymbol{x}^n|\boldsymbol{Y}^n = \boldsymbol{y}^n]$$

最大．

注意到

$$P[\boldsymbol{X}^n = \boldsymbol{x}^n \mid \boldsymbol{Y}^n = \boldsymbol{y}^n] = \frac{P(\boldsymbol{X}^n = \boldsymbol{x}^n, \boldsymbol{Y}^n = \boldsymbol{y}^n)}{P(\boldsymbol{Y}^n = \boldsymbol{y}^n)}.$$

因此，最大后验估计是使分子最大化的 $\boldsymbol{x}^n$ 的值．由式（9.1）可以得到分子的对数如下：

$$\log(\pi_0(x_0)Q(x_0, y_0)) + \sum_{m=1}^{n} \log(P(x_{m-1}, x_m)Q(x_m, y_m)) .$$

定义

$$d(x_0) = -\log(\pi_0(x_0)Q(x_0, y_0)),$$
$$d_m(x_{m-1}, x_m) = -\log(P(x_{m-1}, x_m)Q(x_m, y_m)),$$

那么，最大后验估计就是使下式最小的 $\boldsymbol{x}^n$

$$d(x_0) + \sum_{m=1}^{n} d_m(x_{m-1}, x_m). \tag{9.2}$$

表达式（9.2）可以看作图 9-3 中的路径长度．于是，寻找 MAP 实际等同于解决最短路径问题．最短路径问题已经有一些标准的解法．下面介绍其中一个出名的算法：贝尔曼–福特算法．

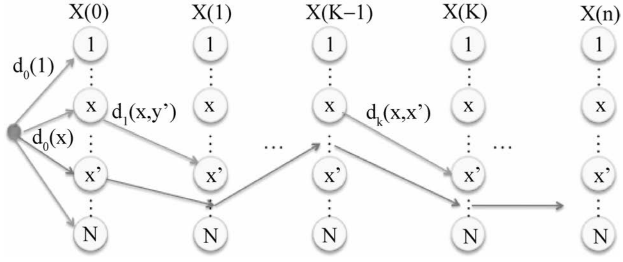


图 9-3 最大后验估计等价于最短路径

对于 $m = 0, \cdots, n$ 和 $x \in \mathcal{X}$, 定义 $V_m(x)$为从 $X(m) = x$ 到列 $X(n)$的最短路径长度. 并且，对于所有 $x \in \mathcal{X}$，令 $V_n(x) = 0$.

$$V_m(x) = \min_{x' \in \mathcal{X}} \{d_{m+1}(x, x') + V_{m+1}(x')\}, x \in \mathcal{X}, m = 0, \cdots, n-1. \quad (9.3)$$

最后，令

$$V = \min_{x \in \mathcal{X}} \{d_0(x) + V_0(x)\}. \quad (9.4)$$

那么，V 就是式（9.2）的最小值.

贝尔曼–福特算法的运行步骤如下.

- 步骤 1：运用式（9.3）递归地计算 $m = n-1, n-2, \cdots, 0$ 时的$\{V_m(x), x \in \mathcal{X}\}$. 在每一步，记录每个节点 x 的外向边中剩余代价为 $V_m(x)$ 的边. 不妨设该边从 $x_m = x$ 出发并指向 $x_{m+1} = s(m, x)$.
- 步骤 2：找到使式（9.4）最小的值 x_0.
- 步骤 3：则 MAP 是序列

$$x_0, x_1 = s(0, x_0), x_2 = s(1, x_1), \cdots, x_n = s(n-1, x_{n-1}).$$

式（9.3）是贝尔曼–福特方程. 该方程是最短路径问题中动态规划方程的一种特殊形式.



图 9-4 理查德 · 贝尔曼（1920—1984）

这里的核心思想是定义从图中每个节点出发到目的地的最短剩余路径，并写出这些量之间的递归方程. 然后，通过从后向前的方式求解动态规划方程就能找到最短的前向路径. 采用最短路径算法求 MAP 的应用叫作维特比算法.

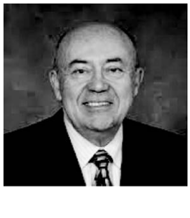


图 9-5 安德鲁 · 维特比（1934— ）

9.3 期望最大化和聚类

C.8

期望最大化（Expectation Maximization，EM）是一类估计分布参数的算法．我们首先通过一个简单的聚类问题来解释这些算法．下一节会将期望最大化应用到隐马尔可夫链的模型中．

聚类问题关注如何将样本点分组到相似值的集群中．我们先用一个简单例子来解释这个问题，然后探讨期望最大化算法．

9.3.1 一个简单的聚类问题

假设你开设了一门概率论课程，想通过 N 个学生的考试成绩$\{X(1), \cdots, X(N)\}$确定哪些学生是 A 等生，哪些是 B 等生．为研究这个问题，假设 A 等生的考试成绩是服从 $\mathcal{N}(a, \sigma^2)$的独立同分布随机变量，而 B 等生的成绩则是服从 $\mathcal{N}(b, \sigma^2)$的独立同分布变量，其中 $a > b$．简单起见，假设 σ^2 的值已知且每个学生有 0.5 的概率是 A 等生．但是，我们不知道参数(a, b).（同样的方法也适用于方差和先验概率都为未知的情况下．）

可以采用以下的启发式方法（见图 9-6）．开始时，猜测(a, b)的值为(a_1, b_1)．如果学生的分数 $X(n) > (a_1+b_1)/2$，将其归为 A 等生，否则将其归为 B 等生．接下来计算 A 等生的平均分 a_2，以及其他学生的平均分 b_2．现在，用(a_2, b_2)替代(a_1, b_1)，重复上述的归类过程．如此反复，直到这两个值看起来收敛为止．这种启发式方法叫作*硬期望最大化算法*（Hard EM）.

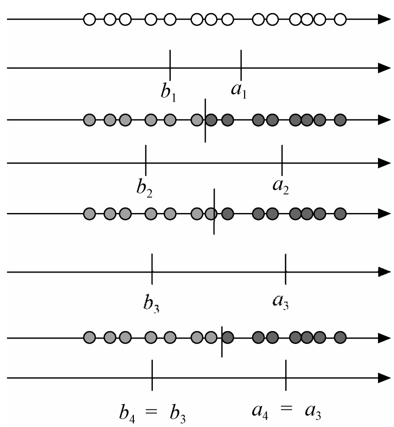


图 9-6 用硬期望最大化算法解聚类问题．初始猜测为(a_1, b_1)，然后得出该情况下最大后验估计与下一个猜测(a_2, b_2)，以此类推

另一个略微不同的启发式方法如下（见图 9-7）．同样以猜测值(a_1, b_1)开始．

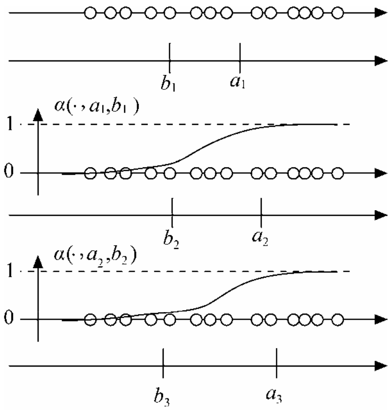


图 9-7　用软期望最大化算法解聚类问题．初始猜测为(a_1, b_1)．然后得出该情况下最大后验估计和下一个猜测(a_2, b_2)，以此类推

运用贝叶斯准则，计算分数为 $X(n)$的学生 n 是 A 等生的概率 $p(n)$．然后，计算下一组估计值

C.8

$$a_2 = \frac{\sum_n X(n)p(n)}{\sum_n p(n)},\ b_2 = \frac{\sum_n X(n)(1-p(n))}{\sum_n (1-p(n))}.$$

用(a_2, b_2)替代(a_1, b_1)后，重复上述过程．因此，在 a_2 的计算过程中，我们考虑了每个学生分数并权衡了该生是 A 等生的概率，b_2 的计算是类似的．

这种启发式方法叫作软期望最大化算法（Soft EM）．

9.3.2　回首再探

在先前的例子中，我们尝试依据观测值 $\boldsymbol{X} = (X_1, \cdots, X_N)$来估计一些参数 $\theta = (a, b)$．令 $\boldsymbol{Z} = (Z_1, \cdots, Z_N)$为学生的属性向量．若学生 n 是 A 等生，则 $Z_n = A$，否则 $Z_n = B$．

我们想要找到使 $f[\boldsymbol{x}|\theta]$最大的 θ 值，以求得 MLE$[\theta|\boldsymbol{X} = \boldsymbol{x}]$．注意，可以把 $f[\boldsymbol{x}|\theta]$表示为

$$f[\boldsymbol{x}|\theta] = \sum_z f[\boldsymbol{x} \mid z, \theta]P[z \mid \theta],$$

其中求和是对总共 2^N 个可能的 $\boldsymbol{Z}$ 值取和．因为 $\boldsymbol{Z}$ 的个数太多，这一步的计算十分复杂．



图 9-8　硬 EM 还是软 EM

硬期望最大化算法实际上是将 $\boldsymbol{z}$ 的和替换成

$$f[\boldsymbol{x}|\boldsymbol{z}^*,\theta]P[\boldsymbol{z}^*|\theta].$$

这里，$\boldsymbol{z}^*$ 是在给定观测值和当前对 θ 的猜测值下最可能的 $\boldsymbol{Z}$ 值．也就是，若当前猜测是 θ_k，那么

$$\boldsymbol{z}^* = MAP[\boldsymbol{Z} \mid \boldsymbol{X}=\boldsymbol{x},\theta_k] = \arg\max_z P[\boldsymbol{Z}=\boldsymbol{z} \mid \boldsymbol{X}=\boldsymbol{x},\theta_k].$$

下一个猜测为

$$\theta_{k+1} = \arg\max_\theta f[\boldsymbol{x} \mid \boldsymbol{z}^*,\theta]P[\boldsymbol{z}^* \mid \theta]$$

软期望最大化算法则采用不同的近似．首先，它用

$$\sum_z \log(f[\boldsymbol{x} \mid \boldsymbol{z},\theta])P[\boldsymbol{z} \mid \theta]$$

代替

$$\log(f[\boldsymbol{x} \mid \theta]) = \log(\sum_z f[\boldsymbol{x} \mid \boldsymbol{z},\theta]P[\boldsymbol{z} \mid \theta]).$$

也就是说，它用对数的期望代替期望的对数．

然后，再将上式转换成

$$\sum_z \log(f[\boldsymbol{x} \mid \boldsymbol{z},\theta])P[\boldsymbol{z} \mid \boldsymbol{x},\theta_k].$$

新的猜测 θ_{k+1} 则是使上式最大的 θ．所以，它把 $\boldsymbol{Z}$ 的分布替代为在给定当前猜测和观测情况下的条件分布．

尽管这是一种启发式算法，但令人惊讶的是，它似乎可以解决一系列的问题．其效果也有一些理论依据．当然，如果这种启发式算法在实际应用中不起作用，就没有人会提及它了．我们可以证明，算法收敛于 $f[\boldsymbol{x}|\theta]$ 的局部最大值．这不是什么特别好的消息，因为大多数问题都不是凸的，因此有许多局部最大值．详见 Roche 的文章“EM algorithm and variants: an informal tutorial”．

9.4 学习：隐马尔可夫链

考虑一个隐马尔可夫链模型．假设(π, P, Q)是待估计参数 θ 的函数，将其记作$(\pi_\theta, P_\theta, Q_\theta)$．我们希望找到一个 θ 值，使得观察到的序列 $\boldsymbol{y}^n$ 在 θ 值下正好是最可能出现的序列．

对于给定的 $\boldsymbol{Y}^n = \boldsymbol{y}^n$，$\theta$ 的 MLE 被定义为

$$MLE[\theta|\boldsymbol{Y}^n = \boldsymbol{y}^n] = \arg\max_\theta P[\boldsymbol{Y}^n = \boldsymbol{y}^n \mid \theta].$$

与聚类例子的推导类似，

$$P[\boldsymbol{Y}^n = \boldsymbol{y}^n \mid \theta] = \sum_{\boldsymbol{x}^n} P[\boldsymbol{Y}^n = \boldsymbol{y}^n \mid \boldsymbol{X}^n = \boldsymbol{x}^n, \theta] P[\boldsymbol{X}^n = \boldsymbol{x}^n \mid \theta]. \tag{9.5}$$

9.4.1 硬期望最大化

硬期望最大化（HEM）算法将对 $\boldsymbol{x}^n$ 的求和换成

$$P[\boldsymbol{Y}^n = \boldsymbol{y}^n \mid \boldsymbol{X}^n = \boldsymbol{x}_*^n, \theta] P[\boldsymbol{X}^n = \boldsymbol{x}_*^n \mid \theta].$$

同时将 $P[\boldsymbol{X}^n = \boldsymbol{x}_*^n \mid \theta]$ 换成

$$P[\boldsymbol{X}^n = \boldsymbol{x}_*^n \mid \boldsymbol{Y}^n, \theta_0],$$

其中，

$$\boldsymbol{x}_*^n = MAP[\boldsymbol{x}^n \mid \boldsymbol{Y}^n, \theta_0].$$

前面提到，$\boldsymbol{x}_*^n$ 的值可以通过维特比算法得到．并且，

$$P[\boldsymbol{Y}^n = \boldsymbol{y}^n \mid \boldsymbol{X}^n = \boldsymbol{x}^n, \theta] = \pi_\theta(x_0) Q_\theta(x_0, y_0) Q_\theta(x_1, y_1) \times \cdots \times P_\theta(x_{n-1}, x_n) Q_\theta(x_n, y_n).$$

9.4.2 训练维特比算法

维特比算法需要知道 P 和 Q．在实际应用中，Q 取决于说话的人，而 P 可能取决于当地方言（比如维利人比伯克利人更常用“喜欢”一词）．如果已经有一个参数化模型，就可以用 HEM 算法．

如果没有参数化模型，可以观测 $\boldsymbol{x}^n$ 和 $\boldsymbol{y}^n$，并采用一个简单的监督训练方法，通过 $\boldsymbol{x}^n$ 和 $\boldsymbol{y}^n$ 中的参数出现频率估计 P 和 Q 的值．比如，可以在 $\boldsymbol{x}^n$ 中用(x_m, x_{m+1})等于(a, b)的次数除以 $x_m = a$ 的次数来估计 $P(a, b)$的值．Q 的值也能用同样的方法估计．

9.5 小结

- 隐马尔可夫链

- ❑ 计算 $MAP[\boldsymbol{X}|\boldsymbol{Y}]$的维特比算法
- ❑ 聚类和期望最大化
- ❑ 隐马尔可夫链的期望最大化

重要方程与公式

隐马尔可夫链的定义	$X(n) = \mathrm{MC}\&P[Y_n\|X_n]$	定义9.1
贝尔曼–福特方程	$V_n(x) = \min_y\{d(x, y)+V_{n+1}(y)\}$	式（9.3）
软期望最大化和硬期望最大化算法	$\theta\to z\to \boldsymbol{x}$；启发式方法计算$MAP[\theta\|\boldsymbol{x}]$	9.3节

9.6　参考资料

Wainwright 和 Jordan 在 *Graphical Models*, *Exponential Families*, *and Variational Inference* 一书中为图模型进行了非常棒的介绍和讲解．它同时涵盖了期望最大化和许多其他有用的技术．

9.7　练习

1．令(X_n, Y_n)为隐马尔可夫链，$Y^n = (Y_0, \cdots, Y_n)$且 $X^n = (X_0, \cdots, X_n)$．维特比算法计算了以下哪项?

- ❑ $MLE[Y^n|X^n]$
- ❑ $MLE[X^n|Y^n]$
- ❑ $MAP[Y^n|X^n]$
- ❑ $MAP[X^n|Y^n]$

2．假设马尔可夫链 X_n 在 $\mathcal{X} = \{a, b\}$中取值．$\pi_0(a) = \pi_0(b) = 0.5$，而且对于 $x\neq x'$有 $P(x, x') = \alpha$，$P(x, x) = 1-\alpha$．又假设 X_n 是通过二元对称信道观测而得，其错误概率为 ε，如图 9-9 所示．应用维特比算法并评估它的性能．

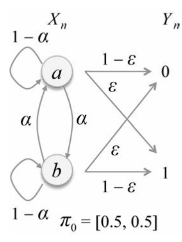


图 9-9　一个简单的隐马尔可夫链

3. 假设一个班级的学生成绩分布服从一个混合高斯分布．其中，服从 $N(\mu_1, \sigma_1^2)$ 分布的概率为 p，服从 $N(\mu_2, \sigma_2^2)$分布的概率为 $1-p$．所有的参数 $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2, p)$都是未知的．

(a) 假设从混合分布中得到了 n 个独立同分布的样本 $y_1, \cdots, y_n$．求 $f(y_1, \cdots, y_n|\theta)$.

(b) 若 $Y_i \sim N(\mu_1, \sigma_1^2)$，令随机变量 X_i 为 0；若 $Y_i \sim N(\mu_2, \sigma_2^2)$，则令其为 1．求 $MAP[X_i|Y_i, \theta]$.

(c) 应用硬期望最大化算法来近似求解 $MLE[\theta|Y_1, \cdots, Y_n]$．你可以使用 Matlab 依照 $\theta = (10, 5, 30, 10, 0.4)$生成 1000 个数据点$(y_1, \cdots, y_{1000})$，然后用你的数据来估计 θ. 你的算法有效吗?

第10章

语音识别—B

主题： 随机梯度，匹配追踪，压缩感知，推荐系统

背景知识： 第 1 章，第 5 章，第 7 章

10.1 在线线性回归

本节解释随机梯度算法．许多学习方案都会用到这一算法．

首先回忆一下，线性回归的目标是寻找参数 a 和 b，使以下的误差最小：

$$\sum_{k=1}^{K}(X_k - a - bY_k)^2,$$

其中(X_k, Y_k)是观测样本，它们是服从同一未知分布 $f_{X,Y}(x, y)$的独立随机变量．

现在假设每次得到新样本就更新参数(a, b)，而不是在等到 K 个样本点之后才计算线性回归．

我们要找到 a 和 b，使下式最小

$$\begin{aligned}
&E((X-a-bY))^2 \\
&\quad = E(X^2)+a^2+b^2E(Y^2)-2aE(X)-2bE(XY)+2abE(Y) \\
&\quad =: h(a,b).
\end{aligned}$$

一种想法是，采用梯度算法最小化 $h(a, b)$．假设在算法的第 k 步，已经计算了$(a(k), b(k))$．此时，梯度算法用与梯度相反的方向更新$(a(k), b(k))$，从而使得 $h(a(k), b(k))$减小．算法更新$(a(k), b(k))$如下：

$$a(k+1) = a(k) - \alpha\frac{\partial}{\partial a}h(a(k),b(k))$$

$$b(k+1) = b(k) - \alpha\frac{\partial}{\partial b}h(a(k),b(k)).$$

其中，α 是一个很小的正数，用以控制更新的步长．因此，

$$a(k+1)=a(k)-\alpha\left[2a(k)-2E(X)+2b(k)E(Y)\right]$$
$$b(k+1)=b(k)-\alpha\left[2b(k)E(Y^2)-2E(XY)+2a(k)E(Y)\right].$$

然而，我们不知道分布，也无法计算期望值．于是用新的样本值代替期望值，即

$$a(k+1)=a(k)-\alpha\left[2a(k)-2X(k+1)+2b(k)Y(k+1)\right]$$
$$b(k+1)=b(k)-\alpha\left[2b(k)Y^2(k+1)-2X(k+1)Y(k+1)+2a(k)Y(k+1)\right].$$

我们使用的是随机梯度算法，而不是梯度算法．这里随机梯度算法可以被视为采用了一个带噪声的梯度来进行更新．直觉告诉我们，如果步长很小，真实梯度与带噪声版本之间的误差平均下来相差无几．

图 10-1 的左图画出了该算法在式（7.4）中的更新值，其中 $\alpha=0.002$，$E(X^2)=1$，$E(Z^2)=0.3$．在该例中，线性最小平方估计是

$$L[X|Y]=a+bY=\frac{1}{1.3}Y=0.77Y.$$

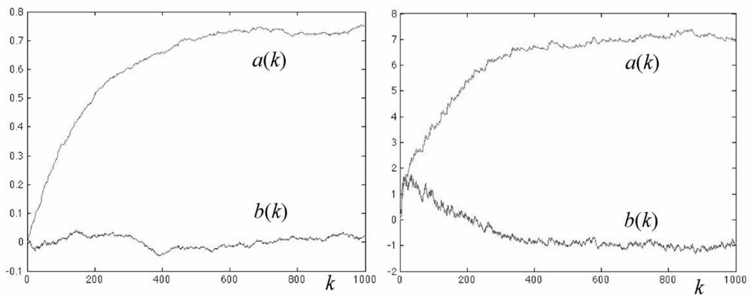


图 10-1　用随机梯度算法“学习”得来的系数，左图为式（7.4），右图为式（7.5）

从图中可以看出，在该算法下，(a_k, b_k)趋近于$(0.77, 0)$．

图 10-1 的右图绘出式（7.5）的系数，$\gamma=0.05$，$\alpha=1$，$\beta=6$．可以看到(a_k, b_k)趋近于$(7, -1)$，正好是线性最小平方估计的值．

10.2　随机梯度投影理论

现在来解释在在线回归的例子中阐述过的随机梯度算法理论．我们先从确定性梯度投影算法开始讨论．

考虑一个定义在凸集合（碗状）上的平滑凸函数．最小化该函数（寻找碗底）的标准算法就是梯度投影算法．这个算法类似于沿着最陡峭的斜坡，用越来越小的步子一步步跳下山．投影可以确保算法给出的答案总在可接受的集合上，而算法的步长随着时间减小保证了不会跳过最小值．

随机梯度投影算法与梯度投影算法类似，区别在于随机梯度投影算法采用带噪声的梯度．随着步长越来越小，梯度的误差平均下来所剩无几，因此算法最终收敛到函数的最小值．

下面先来复习一下梯度投影算法，然后讨论随机梯度投影算法．

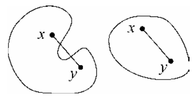


图 10-2　一个非凸集（左）和一个凸集（右）

10.2.1　梯度投影

假设要求可导凸函数 $f(x)$ 在 $\Re^d$ 的一个闭凸子集 $\mathcal{C}$ 上的最小值．根据定义，$\mathcal{C}$ 是凸集合意味着

$$\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y} \in \mathcal{C}, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{C} \text{ 且 } \theta \in (0,1). \tag{10.1}$$

即 $\mathcal{C}$ 包含了任意两点间的一条线段，在集合边界上没有洞或者扭结．

另外，回顾凸函数的定义：如果

$$f(\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y}) \leqslant \theta f(\boldsymbol{x}) + (1-\theta) f(\boldsymbol{y}), \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{C} \text{ 且 } \theta \in (0,1), \tag{10.2}$$

则函数 $f: \mathcal{C} \to \Re$ 是一个凸函数．

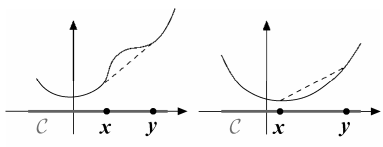


图 10-3　一个非凸的函数（左）和一个凸函数（右）

该问题的标准的解法是梯度投影算法（GP）：

$$\boldsymbol{x}_{n+1} = [\boldsymbol{x}_n - \alpha_n \nabla f(\boldsymbol{x}_n)]_C \ (n \geqslant 0).$$

这里，

$$\nabla f(\boldsymbol{x}) := [\frac{\partial}{\partial x_1} f(\boldsymbol{x}), \cdots, \frac{\partial}{\partial x_d} f(\boldsymbol{x})]'$$

是函数 $f(\cdot)$ 在 $\boldsymbol{x}$ 处的梯度，而 $[\boldsymbol{y}]_C$ 代表集合 $\mathcal{C}$ 中最接近 $\boldsymbol{y}$ 的点，也叫作 y 在 $\mathcal{C}$ 上的投影．常数 $\alpha_n > 0$ 是算法的步长．

举一个简单的例子，令 $f(x) = 6(x-0.2)^2$，且 $x \in \mathcal{C} := [0, 1]$．这里的系数 6 只是为了说明开始时算法需要大的步长，并且需要通过投影保证答案的合理性．下面采用 $\alpha_n = 1/n$，$x_0 = 0$．由此，算法如下：

$$x_{n+1} = \left[x_n - \frac{12}{n}(x_n - 0.2) \right]_{\mathcal{C}}. \quad (10.3)$$

式（10.3）等价为：

$$y_{n+1} = x_n - \frac{12}{n}(x_n - 0.2) \quad (10.4)$$

$$x_{n+1} = \max\{0, \min\{1, y_{n+1}\}\}. \quad (10.5)$$

其中，$y_0 = x_0$.

如图 10-4 所示，当步长较大时，更新值 y_{n+1} 可能落在集合 $\mathcal{C}$ 外并被投影回这个集合．最终，更新值 x_n 落在集合 $\mathcal{C}$ 内.

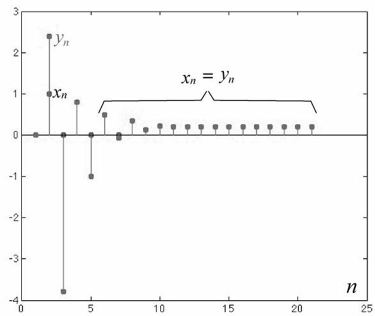


图 10-4　梯度投影算法式（10.4）~（10.5）（另见彩插）

有许多能保证算法收敛到 $f(\cdot)$在 $\mathcal{C}$ 上唯一最小值点的充分条件，下面就是一例.

定理 10.1　梯度投影算法的收敛性

设 $f(x)$为凸集合 $\mathcal{C}$ 上可导的凸函数，且满足

$$f(x)\text{在 } \mathcal{C} \text{ 上有唯一的最小值点 } x^* \quad (10.6)$$

$$\|\nabla f(x)\|^2 \leqslant K, \forall x \in \mathcal{C} \quad (10.7)$$

$$\sum_n \alpha_n = \infty \text{ 且} \sum_n \alpha_n^2 < \infty. \quad (10.8)$$

则

$$\text{当 } n \to \infty \text{时，} x_n \to x^*. \qquad \square$$

证明

证明的思想如下．令 $d_n = \frac{1}{2}\|x_n - x^*\|^2$，并固定一个小常量 $\varepsilon > 0$．可以证明存在某 $n_0(\varepsilon)$，使得 $n \geqslant n_0(\varepsilon)$时，

$$d_{n+1} \leqslant d_n - \gamma_n，当 d_n \geqslant \varepsilon 时 \tag{10.9}$$

$$d_{n+1} \leqslant 2\varepsilon，\quad 当 d_n < \varepsilon 时. \tag{10.10}$$

而且，在式（10.9）中，$\gamma_n > 0$ 且 $\sum_n \gamma_n = \infty$.

从式（10.9）可以看到，存在 $n = n_1(\varepsilon) \geqslant n_0(\varepsilon)$ 使得 $d_n < \varepsilon$. 再由式（10.9）和式（10.10）可知，对所有 $n \geqslant n_1(\varepsilon)$，都有 $d_n < 2\varepsilon$. 由于以上论证对任意 $\varepsilon > 0$ 均成立，得到 $x_n \to x^*$.

为了证明式（10.9）和式（10.10），首先要说明

$$d_{n+1} \leqslant d_n + \alpha_n \left(x^* - x_n\right)^T \nabla f\left(x_n\right) + \frac{1}{2}\alpha_n^2 K. \tag{10.11}$$

式（10.11）可由下面的步骤得到：

$$\begin{aligned} d_{n+1} &= \frac{1}{2}\left\|\left[x_n - \alpha_n \nabla f\left(x_n\right)\right]_{\mathcal{C}} - x^*\right\|^2 \\ &\leqslant \frac{1}{2}\left\|x_n - \alpha_n \nabla f\left(x_n\right) - x^*\right\|^2 \end{aligned} \tag{10.12}$$

$$\leqslant d_n + \alpha_n \left(x^* - x_n\right)^T \nabla f\left(x_n\right) + \frac{1}{2}\alpha_n^2 K. \tag{10.13}$$

不等式（10.12）根据凸集合上投影的非扩张性得到，即

$$\left\|[x - y]_{\mathcal{C}}\right\| \leqslant \|x - y\|.$$

这个性质从图 10-5 中可以清晰地看出，并且不难证明.

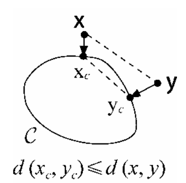


图 10-5 在凸集合上的投影是非扩张的

由于 $\sum_n \alpha_n^2 < \infty$，所以 $\alpha_n \to 0$. 将这一性质用到式（10.13）和式（10.7）中即可推出式（10.10）.

最后只需要证明式（10.9）. 如图 10-6 所示，由 $f(\cdot)$ 的凸性可知

$$(x^* - x)^T \nabla f(x) \leqslant f(x^*) - f(x). \tag{10.14}$$

而且，若 $d_n \geqslant \varepsilon$，则存在 $\delta(\varepsilon) > 0$，使得 $f(x^*) - f(x_n) \leqslant -\delta(\varepsilon)$. 因此，只要 $d_n \geqslant \varepsilon$，就有

$$(x^* - x_n)^T \nabla f(x_n) \leqslant -\delta(\varepsilon).$$

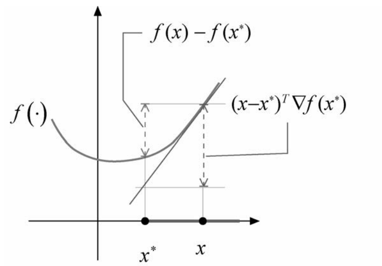


图 10-6　不等式（10.14）（另见彩插）

结合上式与式（10.11），

$$d_{n+1} \leqslant d_n - \alpha_n \delta(\varepsilon) + \frac{1}{2}\alpha_n^2 K.$$

现在，令

$$\gamma_n = \alpha_n \delta(\varepsilon) - \frac{1}{2}\alpha_n^2 K. \tag{10.15}$$

因为 $\alpha_n \to 0$，所以当 $n \geqslant n_2(\varepsilon)$时，必然存在 $n_2(\varepsilon)$ 使得 $\gamma_n > 0$．而且，由式（10.8）可知 $\sum_n \gamma_n = \infty$．现在用 $\max\{n_0(\varepsilon), n_2(\varepsilon)\}$替换 $n_0(\varepsilon)$，式（10.9）即得证．♣

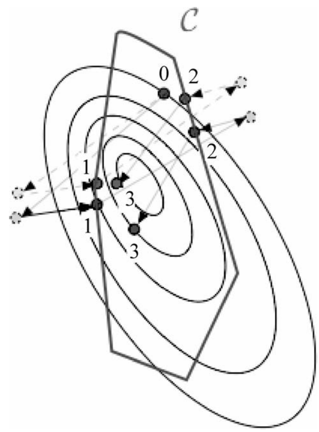


图 10-7　该图展示了$f(\cdot)$的等位线和凸集合 $\mathcal{C}$，以及开始的几次迭代．梯度投影算法用红色表示，随机梯度投影算法用蓝色表示（另见彩插）

10.2.2　随机梯度投影算法

在很多情况下，无法直接计算函数的梯度$\nabla f(\boldsymbol{x}_n)$．不过通常可以得到这个梯度的随机估计值$\nabla f(\boldsymbol{x}_n)+\eta_n$，其中$\eta_n$是一个随机变量．如果有一个很小的$\eta_n$，那么当用$\nabla f(x_n)+\eta_n$替代$\nabla f(\boldsymbol{x}_n)$时，梯度投影算法仍会收敛于 x^*．本节的重点就是证明这一点．

随机梯度算法如下：

$$\boldsymbol{x}_{n+1}=[\boldsymbol{x}_n-\alpha_n\boldsymbol{g}_n]_{\mathcal{C}}, \tag{10.16}$$

其中，

$$\boldsymbol{g}_n=\nabla f(\boldsymbol{x}_n)+\boldsymbol{z}_n+\boldsymbol{b}_n \tag{10.17}$$

是真实梯度带噪声的估计．在式（10.17）中，z_n是一个平均值为 0 的随机变量，用于对估计的噪声进行建模；b_n是一个常量，用于对梯度估计的偏差进行建模．

举一个简单的例子，设 $f(x)=6(x-0.2)^2$，且 $x\in\mathcal{C}:=[0,1]$，并将常数设置为 $\alpha_n=1/n$，$b_n=0$ 以及 $x_0=0$，这时候，算法的更新步骤如下：

$$x_{n+1}=[x_n-\frac{12}{n}(x_n-0.2+z_n)]_{\mathcal{C}}. \tag{10.18}$$

在这个表达式中，z_n是服从 $U[-0.5,\ 0.5]$的独立同分布随机变量．图 10-8 展示了该算法生成的值．可以看到，尽管收敛速度很慢，但这一方法最终收敛到了函数的最小值点 $x^*=0.2$．

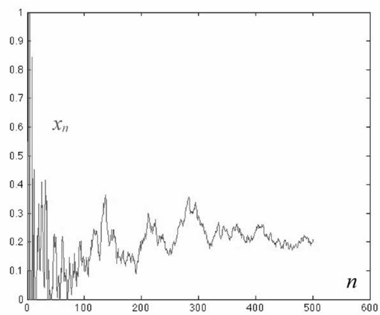


图 10-8　随机梯度投影算法式（10.18）

为了让算法（10.16）~（10.17）收敛，噪声估计值 z_n和偏置 b_n需要很小．下面的定理给出了具体的要求．

定理 10.2　随机梯度投影的收敛性

假设 $\mathcal{C}$ 有界且满足

$$f(\cdot)\text{在 } \mathcal{C} \text{ 上有唯一的最小值点 } \boldsymbol{x}^*; \tag{10.19}$$

$$\left\|\nabla f\left(\boldsymbol{x}\right)\right\|^2\leqslant K, \forall x\in\mathcal{C}; \tag{10.20}$$

$$\alpha_n>0, \sum_n\alpha_n=\infty,\ \sum_n\alpha_n^2<\infty. \tag{10.21}$$

除此之外，假设

$$\sum_{n=0}^{\infty}\alpha_n\|\boldsymbol{b}_n\|<\infty; \tag{10.22}$$

$$E\left[z_{n+1}\middle|z_0,z_1,\cdots,z_n\right]=0; \tag{10.23}$$

$$E\left(\|z_n\|^2\right)\leqslant A, n\geqslant 0. \tag{10.24}$$

那么，$\boldsymbol{x}_n\to\boldsymbol{x}^*$的概率为 1. □

证明

该证明本质上和梯度投影的证明一致.

不等式（10.11）变成

$$d_{n+1}\leqslant d_n+\alpha_n\left(x^*-\boldsymbol{x}_n\right)^T\left[\nabla f\left(\boldsymbol{x}_n\right)+\boldsymbol{z}_n+\boldsymbol{b}_n\right]+\frac{1}{2}\alpha_n^2K. \tag{10.25}$$

相应地，式（10.15）中的 γ_n 变成

$$\gamma_n=\alpha_n\left[\delta\left(\varepsilon\right)+\left(\boldsymbol{x}^*-\boldsymbol{x}_n\right)^T\left(\boldsymbol{z}_n+\boldsymbol{b}_n\right)\right]-\frac{1}{2}\alpha_n^2K. \tag{10.26}$$

现在，式（10.23）意味着 $\boldsymbol{v}_n:=\sum_{m=0}^{n}\alpha_m\boldsymbol{z}_m$ 是鞅[①]. 由式（10.24）和式（10.21）可知，对于所有 n，$E\left(\|\boldsymbol{v}_n\|^2\right)\leqslant A<\infty$. 由鞅收敛定理 10.3 可得，$v_n$ 收敛到一个有限的随机变量. 将其与式（10.22）相结合可得[②] $\sum_{m=n}^{\infty}\alpha_m\left[\boldsymbol{z}_m+\boldsymbol{b}_m\right]\to 0$. 由于$\left\|\boldsymbol{x}_n-\boldsymbol{x}^*\right\|$有界，误差的估计值随着 n 的增大可以忽略. 剩下的证明步骤与梯度投影的证明步骤相同.

♣

10.2.3 鞅收敛定理

在 13.8 节，我们会讨论鞅的理论. 这里仅讨论证明定理 10.2 所需要的一些概念.

令$\{x_n, y_n, n\geqslant 0\}$为随机变量，且对于所有 n，x_n 的期望值 $E(x_n)$均存在. 如果

$$E\left[x_{n+1}\middle|x_m,y_m,m\leqslant n\right]=x_n,\forall n,$$

则序列 x_n 被称为关于$\{(x_m, y_m), m\geqslant 0\}$的鞅.

定理 10.3 鞅收敛定理

如果鞅 x_n 使得对于所有的 n 有 $E(x_n^2)\leqslant B<\infty$，那么它以概率 1 收敛于一个有限的随机变量. □

---

① 见下一小节.

② 回顾：若级数 $\sum_n w_n$ 收敛，则当 $n\to\infty$时，级数的尾端 $\sum_{m\geqslant n}w_m$ 收敛于 0.

证明详见定理 13.17.

10.3 大数据

网络的出现使得许多领域中的数据收集问题不再复杂. 例如，现在我们能很容易地找到人们喜爱的书籍、电影、餐厅，经常浏览的网站，移动习惯，医疗记录，以及传感器的测量值. 这些数据在人们进行选择的时候往往能起到很好的辅助效果. 比如，推荐人们可能喜爱的商品，提供可能有效的治疗方法，引荐可能想与之交谈的人，调查谁在和谁聊天，找寻有效的管理技巧，等等. 不仅如此，与数据收集同步发展的存储技术、数据库和云计算等科技也让实时处理这些数据并完成目标成为了可能. 本节讨论这个方向上的一些问题和解决这些问题的算法.

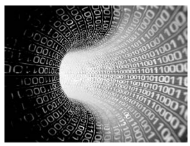


图 10-9 网络提供了获得大量数据的接口，该如何从数据中提取有用的信息呢

10.3.1 相关数据

一件事情的结果往往取决于很多因素，但是哪些因素是最关键的呢？比如，一名学生在大学的成功通常取决于高中成绩、大学先修课程的分数和高考分数. 如何找到能准确预测其成功的因素呢？类似的情形还有预测患上某种疾病的概率，预测某种疗法的成功率，等等.

确定这些重要因素对于改进结果大有裨益. 比如，如果知道一名学生在大学成功与否主要由其高中时期的读书量及其玩电脑游戏的时间所决定，那么就可以提出合适的建议，提高学生拥有成功大学学习经历的机会.

这一问题可以规范成以下的数学问题：设事件的结果为 Y. 已知其与一系列因素有关，因此用一个有 $N \gg 1$ 个分量的向量 $\boldsymbol{X}$ 表示这些因素. 比如，若 Y 是大学四年结束后的 GPA，$\boldsymbol{X}$ 的第一个分量 X_1 可能指的是高中的 GPA，第二个分量 X_2 是高考成绩，X_3 是学生撰写读书报告的书籍数目，等等. 尽管 $N \gg 1$，但是通常其中只有很少的 $\boldsymbol{X}$ 分量真正对结果 Y 有实质性影响. 不过在这个问题中，我们不预先假定已知这些分量.

假设我们想在 $\boldsymbol{X}$ 的 6 个分量的基础上预测 Y，怎么知道需要考虑哪些分量呢？考虑到 $\mathcal{N} = \{1, 2, \cdots, N\}$ 有许多 6 个元素的子集（大约有 $N^6/6!$ 个），这个问题其实难度不小. 尤其是当 N 很大时，这一问题的多种组合令它很棘手. 为取得进展，我们将这

一问题稍作改变，并求助于启发式方法[①].

将问题转变成寻找 $\boldsymbol{b}=(b_1, \cdots, b_N)$使得下式最小化：

$$J(\boldsymbol{b})=E((Y-\sum_n b_n X_n)^2).$$

其中，$\boldsymbol{b}$ 的代价函数不能大于一定的规定值：

$$C(\boldsymbol{b})=\sum_n |b_n|.$$

这叫作 LASSO（Least Absolute Shrinkage and Selection Operator，最小绝对收缩和选择算子）问题，亦称套索问题．这里用系数的代价替换了对分量数量的限制．不难看到，变换后的问题与之前的问题很类似．并且对系数代价的限制保证了问题的解会有很多 b_n 等于 0．这也很直观：如果某一分量不如其他分量重要，则其在解中的系数很可能等于 0．

下面来解释这个问题的含义．为了简化运算，假设 Y 和 $\boldsymbol{X}$ 的期望值为零，并且

$$Y=\sum_n B_n X_n+Z,$$

其中 Z 为 $\mathcal{N}(0, \sigma^2)$变量，而系数 B_n 为独立并且分布为

$$f_B(b)=\frac{\lambda}{2}\exp\{-\lambda|b|\}$$

的随机变量．

在这种情况下，

$$\begin{aligned}
MAP[\boldsymbol{B}|\boldsymbol{X}=\boldsymbol{x}, Y=y] &= \arg\max_b f_{\boldsymbol{B}|\boldsymbol{X},Y}[\boldsymbol{b}|\boldsymbol{x}, y] \\
&= \arg\max_{\boldsymbol{b}} f_{\boldsymbol{B}}(\boldsymbol{b}) f_{Y|\boldsymbol{X}}[y|\boldsymbol{x}] \\
&= \arg\max_{\boldsymbol{b}} \exp\left\{-\frac{1}{2\sigma^2}(y-\sum_n b_n x_n)^2\right\}\exp\left\{-\lambda\sum_n |b_n|\right\} \\
&= \arg\min_{\boldsymbol{b}}\left\{(y-\sum_n b_n x_n)^2+\mu\sum_n |b_n|\right\}.
\end{aligned}$$

其中 $\mu=2\lambda\sigma^2$．这个问题的形式是 LASSO 问题的拉格朗日乘子规范形式，其中对系数代价 $C(\boldsymbol{b})$的限制改由在目标函数中加入惩罚项 $\mu C(\boldsymbol{b})$代替．因此，LASSO 问题等同于在上述随机条件假设下求 $MAP[B|X, Y]$的问题．

现在来看一个贪心算法．这个算法每次选择一个分量，且每次的选择都尽量地最大化成效．首先假设可以只在 X 的 N 个分量中选择一个分量 X_n，那么

$$L[Y|X_n]=\frac{\text{cov}(Y, X_n)}{\text{var}(X_n)}X_n =: b_n X_n$$

① 如果你砸不开一个坚果，就再找另一个试试．（工程和数学的区别？）

且

$$E\left(\left(Y-L[Y|X_n]\right)^2\right)=\text{var}(Y)-\frac{\text{cov}(Y,X_n)^2}{\text{var}(X_n)}.$$
$$=\text{var}(Y)-|\text{cov}(Y,X_n)|\times|b_n|.$$

可以将其理解为每一个单位的系数代价 $C(b_n)=|b_n|$带来了目标 $J(b_n)$的减量$|\text{cov}(Y, X_n)|$．这么一来，每次选择使得“每单位代价奖励”$|\text{cov}(Y, X_n)|$最大的分量就显得理所当然了．不妨设这一分量为 X_1，并令 $\hat{Y}_1=L\left[Y|X_1\right]$．

第二步，在保持 X_1 不变的情况下，寻找第二个分量 X_n，$n\neq 1$．此时，

$$E\left(\left(Y-b_1X_1-b_nX_n\right)^2\right)$$
$$=E\left(\left(Y-b_1X_1\right)^2\right)-2b_n\,\text{cov}\left(Y-b_1X_1,X_n\right)+b_n^2\,\text{var}\left(X_n\right).$$

该表达式在

$$b_n=\frac{\text{cov}\left(Y-b_1X_1,X_n\right)}{\text{var}\left(X_n\right)}$$

时最小，且其最小值等于

$$E\left(\left(Y-b_1X_1\right)^2\right)-\frac{\text{cov}\left(Y-b_1X_1,X_n\right)^2}{\text{var}\left(X_n\right)}.$$

与第一步选择时一样，将总的系数代价 $C(b_1, b_n)$投资在 b_n 上会带来 $J(b_1, b_n)$的单位目标减量

$$\left|\text{cov}\left(Y-b_1X_1,\ X_n\right)\right|.$$

这说明，第二步要选择的 X_n 应该是与 $Y–b_1X_1$ 有最大协方差的 X_n.

由此得到下面的算法，叫作*逐步回归算法*．在第 k 步，算法找到与残余误差 $Y-\hat{Y}_k$ 最相关的分量 X_n，其中 $\hat{Y}_k$ 是当前的估计．具体来说，算法如下所示．

- 第 0 步：$\hat{Y}_0=E\left(Y\right)$ 且 $S_0=\varnothing$.
- 第 k+1 步：找到 $n\notin S_k$，使得 $E((Y-\hat{Y}_k)X_n)$ 最大，

  令 $S_{k+1}=S_k\cup\{n\}$，$Y_{k+1}=L[Y|X_n, n\in S_{k+1}]$, $k=k+1$.
- 重复上述步骤，直到 $E((Y-\hat{Y}_k)^2)\leqslant\varepsilon$.

实际问题通常会包含一系列的结果$\{Y^m,\ m=1,\ \cdots,\ M\}$以及相对应的因素 $X^m=(X_1^m, X_2^m, \cdots, X_N^m)$．在这里，每一个 m 对应一个样本，比如说一个学生在大学成功的样本．根据这些样本，可以通过样本平均估计真实的期望值．因此，如果在第 k 步已经计算出用以估计

$$\hat{Y}_k^m = b_1 X_1^m + \cdots + b_k X_k^m .$$

的系数(b_1, ⋯, b_k). 接下来就可以通过

$$\frac{1}{M}\sum_{m=1}^{M}(Y^m - \hat{Y}_k^m)X_n^m$$

估计 $E((Y-\hat{Y}_k)X_n)$. 这里同样可以用线性回归近似 $L[Y|X_n, n \in S_{k+1}]$.

虽然可能有数百万名学生的数据，但是根据大数定律，用于估计均值和协方差的样本数 M 不一定要非常大. 因此，用几千个样本就能获得比较合理的估值. 这里也可以用前面介绍过的样本矩来计算这些估计的置信区间.

信号处理中有一个类似的算法叫作*匹配追踪*，由 Mallat 和 Zhang 于 1993 年在“Matching Pursuits with Time-Frequency Dictionaries”一文中提出. 该问题考虑的场景是寻找一个信号的紧凑表示（如图像或语音）. 需要将信号的表示考虑为一些基函数的线性组合，而匹配追踪算法则能找出一组最重要的基函数.

一个例子

这里用一个非常简单的例子来说明每一个步骤. 假设所有例子中随机变量的期望值均 0. 我们有 3 个样本，相应的协方差矩阵为

$$\sum_Z = \begin{bmatrix} 4 & 3 & 2 & 2 \\ 3 & 4 & 2 & 2 \\ 2 & 2 & 4 & 1 \\ 2 & 2 & 1 & 4 \end{bmatrix},$$

其中 $\boldsymbol{Z}' = (Y, X_1, X_2, X_3) = (Y, \boldsymbol{X}')$.

首先尝试逐步回归法. 因为与 Y 最相关的分量为 X_1，所以

$$\hat{Y}_1 = L\left[Y\middle|X_1\right] = \frac{\text{cov}(Y, X_1)}{\text{var}(X_1)}X_1 = \frac{3}{4}X_1 =: b_1 X_1.$$

下一步计算 $E(X_n(Y-\hat{Y}_1))$ 的相关性.

$$E(X_2(Y-\hat{Y}_1)) = E(X_2(Y-b_1X_1)) = 3-2b_1 = 1.5$$
$$E(X_3(Y-\hat{Y}_1)) = E(X_3(Y-b_1X_1)) = 2-b_1 = 0.5.$$

现在算法选择 X_2 为下一个分量.

$$\hat{Y}_2 = L[Y|X_1, X_2] = [3 \quad 2]\begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}^{-1}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \frac{2}{3}X_1 + \frac{1}{6}X_2.$$

可以计算误差的方差为

$$E((Y-\hat{Y}_2)^2) = \frac{5}{3}.$$

10.3.2　压缩感知

通常来说，看起来复杂的事物可能蕴藏着非常简单的结构．比如，图 10-10 中所示的信号 $s(t)$就是 3 个正弦函数的和：

$$s(t)=\sum_{i=1}^{3}b_i\sin(2\pi\phi_i t),\ \ t\geqslant 0. \tag{10.27}$$

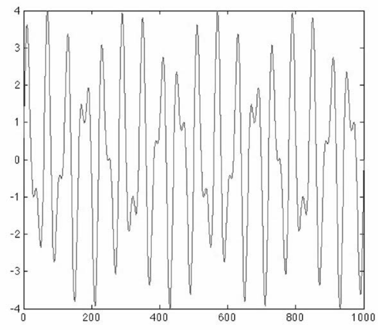


图 10-10　看起来复杂的信号实际上是 3 个正弦函数的和

在信号处理中有一个经典结论，叫作奈奎斯特采样定理．它指出，如果以信号最大频率的 2 倍速率采样，即采样间隔时间 T 满足在 $1/T$ 至少是信号最大频率的两倍，就可以通过采样值准确地恢复一个信号．根据这一结论，只要对于所有 $i=1, 2, 3$，都有 $T<1/\left(2\phi_i\right)$，就可以通过获取每 T 秒的值来重构信号 $s(t)$．但是，对于（10.27），我们可以通过 6 个参数$\left\{b_i,\phi_i,i=1,2,3\right\}$的值来完全描述 $s(t)$．在这个例子中，比较明显的一点是我们不需要知道很多时刻 t_k 的样本值 $s(t_k)$才能重建这 6 个参数以及 $t\geqslant 0$ 时的信号 $s(t)$．同时，如果采样的时间 t_k 为随机选取的，那么重构的信号应当是唯一的．同样的结论在使用不同的基函数时也成立（如多项式函数或小波函数）．

可以看到，如果一个信号可以通过基函数（如正弦函数）简单地表达出来，那么就可能采用少量样本完全重构这一信号．知道这种可能性的存在似乎并没有减小问题的困难．可是，如果稍微变换一下角度，这个问题其实是可以解决的．假设有下面一组函数：

$$g_n\left(t\right)=\sin\left(2\pi f_n t\right), t\geqslant 0, n=1,\cdots,N.$$

再假设 $s(t)$中的频率$\left\{\phi_1,\phi_2,\phi_3\right\}$只在频率$\{f_n, n=1, \cdots, N\}$中取值．可以尝试找到一个向量$\boldsymbol{a}=\left\{a_n,n=1,\cdots,N\right\}$使得对于 $k=1, \cdots, K$，

$$s\left(t_k\right)=\sum_{n=1}^{N}a_n g_n\left(t_k\right).$$

只有 3 个函数的时候，可以通过选择合适的函数来解决这一问题．但是，应当如何系统化地做这件事呢？



第一个想法是将问题规范成下面的问题： ♣

$$\text{最小化} \sum_n 1(a_n \neq 0) \text{ 使得对于 } k=1,\ \cdots, K,\quad \text{有 } s(t_k) = \sum_n a_n g_n(t_k)\ .$$

即用最少的基函数线性组合来表示 $s(t)$.

遗憾的是，这个问题和前一节的问题一样，需要考虑太多非零系数 a_n 的选择组合．因此这一方法也十分困难．不过，我们可以同样将这个问题转化为一个类似的问题：

$$\text{最小化} \sum_n |a_n| \text{ 使得对于 } k=1,\ \cdots, K \text{ 有 } s(t_k) = \sum_n a_n g_n(t_k)\ . \qquad (10.28)$$

最小化系数 a_n 的绝对值之和是对最小化非零系数个数的一种放缩．（从简单的例子中可以看出，选择用 $\sum_n |a_n|^2$ 而不是 $\sum_n |a_n|$ 进行近似通常会导致恢复失败．）在这种情况下，当 K 很大的时候，结果正确的概率很大．

定理 10.4　基于随机样本的准确恢复 [Candes & Romberg，2007]

若

$$K \geqslant C \times B \times \log(N),$$

则信号 $s(t)$ 可以通过求解采用 K 个采样值的式（10.28）以很高的概率准确恢复．在表达式中，C 是很小的常量，B 是构成 $s(t)$ 的正弦函数个数，而 N 为正弦函数的个数． □

注意到，定理 10.4 是一个概率性的结果．实际上，我们的确有可能不巧地遇上所有采样值都为 0 的情况（见图 10-10）．在这种情况下，我们得到的采样点无法被用来恢复 $s(t)$，也就是采样点无法用于完全恢复信号．因此，定理 10.4 表明这种情况发生的概率很小．

在我们的例子中，$B = 3$．所以如果 $N \leqslant 100$，就可以期望从 $3\log(100) \approx 14$ 个样本中准确恢复信号 $s(t)$.

式（10.28）的问题与下述线性规划问题等价．这意味着我们能高效地解决它：

$$\text{最小化} \sum_n b_n \text{ 使得对于 } k = 1, \cdots, K,\ \text{有 } s(t_k) = \sum_n a_n g_n(t_k)\ ,$$

$$\text{对于 } n = 1,\ \cdots, N,\ \text{有} -b_n \leqslant a_n \leqslant b_n. \qquad (10.29)$$

1．数值例

假设

C.10

$$s(t) = \sin(2\pi t) + 2\sin(2.4\pi t) + 3\sin(3.2\pi t),\quad t \in [0,1]. \qquad (10.30)$$

这里可以看到 $s(t)$ 中的频率是 $\phi_1 = 1, \phi_2 = 1.2, \phi_3 = 1.6$ ．函数集合为

$$\{g_n(t) = \sin(2\pi f_n t), n = 1, \cdots, 100\},$$

其中 $f_n = n/10$.

在该集合中，正弦函数的频率为 0.1, 0.2, ⋯, 10．所以，$s(t)$中的频率包含在集合中，且最佳恢复为

$$s(t) = \sum_n a_n g_n(t),$$

其中 $a_{10} = 1$，$a_{12} = 2$，$a_{16} = 3$，而所有其他系数为 0．前面的推导告诉我们，大概需要 14 个样本点即可恢复原信号．因此，我们随机选取了 15 个在[0, 1]间均匀分布的采样时刻 t_k，然后运用 Matlab 求解式（10.29）．图 10-11 展示了我们的结果．代码在附录 C.10 中．

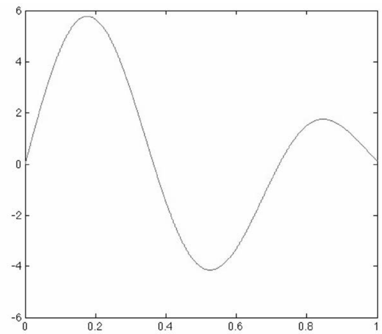


图 10-11　通过在[0, 1]间均匀选取 15 个样本点，完全恢复信号（10.30）．在图里，原始信号与恢复重建的信号重合

2. 另一个例子

图 10-12 摘自 Candes 和 Romberg 的论文"Sparsity and Incoherence in Compressive Sampling"，展示了另一个例子．左边的原始图片有大概 100 万像素，但是，它能被 25 000 个名为小波的函数线性组合而成．所以，压缩感知告诉我们，该图片可由几倍于 25 000 个随机采样的像素点完全恢复．结果也确实如此，可以用约 96 000 个像素点重建该图．



图 10-12　有 10^6 像素点的原图（左）和由 96 000 个随机选取的像素点重构的图像（右）

10.3.3 推荐系统

这里考虑的是如何知道用户喜欢什么样的电影．这个问题的一种规范描述如下：有一个 $K \times N$ 的矩阵 Y，矩阵的元素 $Y(k, n)$代表用户 k 有多喜爱电影 n．但是，我们通常看不到整个矩阵，只能通过用户看过的电影以及他们对电影的评价来获得很有限的几个样本值．我们要考虑的问题是把这个矩阵补全．

矩阵填充这一想法源于矩阵元素并不独立．比如，假设甲和乙看过同样的 5 部电影，给出的评价相当．这时候，甲又看了一部电影并且觉得很喜欢．那么乙也很有可能会喜欢这部电影．

为了规范地描述矩阵元素间的依赖性，我们观察到，虽然有成千上万部电影，但是影响用户喜爱程度的因素往往只有几个．因此可以合理地认为，矩阵的许多列是一些公共向量的组合，而这些向量对应着影响用户排序的因素．所以，这几个独立向量的线性组合构成了这些列．因此，矩阵 Y 只有少量线性无关的列，即 Y 为一个低秩矩阵[①]．那么，我们能否根据观测到的元素来恢复一个低秩的矩阵 Y?

首先规范化这一问题：

$$\text{最小化 } \operatorname{rank}(X)\text{，且保证 } X(k,n) = M(k,n), \forall (k,n) \in \Omega.$$

这里，$\{M(k,n),(k,n) \in \Omega\}$ 是矩阵已知元素的集合．我们希望找到一个秩最低的矩阵 X 与已知的观察值相一致．

如前所述，这样的问题比较难．为了进行简化，我们用核范数$\|X\|_*$代替秩，

$$\|X\|_* = \sum_i \sigma_i \ ,$$

σ_i 为矩阵 X 的特征值（见附录 B.6）．这里非零特征值的个数由矩阵的秩决定．因为核范数是矩阵元素的凸函数，所以很容易求解这一凸优化问题．值得一提的是，在压缩感知问题中，通过改进问题求得的解往往接近最优．

定理 10.5　由随机元素完整地恢复矩阵 [Candes & Recht，2009]

若观察到的元素是随机均匀选取的，且至少有

$$Cn^{1.25} r \log(n)$$

个观测点，则问题

$$\text{最小化 } \|X\|_* \text{ 使得 } X(k,n) = M(k,n), \forall (k,n) \in \Omega$$

的解以很大概率等于矩阵 Y. 该式中，C 是一个较小的常量，$n = \max\{K, N\}$，

---

① 矩阵的秩是矩阵中线性无关列的数目．

而 r 是矩阵 Y 的秩. □

这一结论在观测点个数远小于 $K \times N$（Y 的元素个数）的情况下十分有用. 本章后面的参考资料中有很多该问题的拓展和有关数值求解的细节问题.

10.4　小结

- 在线线性回归
- 凸集合和函数
- 梯度投影算法
- 随机梯度投影算法
- 鞅收敛定理
- 大数据：相关数据，压缩感知，推荐系统

重要方程与公式：

凸集合	如果包含它的任意弦	式（10.1）
凸函数	如果在它的切线以上	式（10.2）
梯度投影的收敛	如果有唯一的最小值点和有界的梯度	定理10.1
随机梯度投影的收敛	如果界限不定且噪声有变化	定理10.2
鞅收敛定理	在L^1或L^2上有界的鞅以概率1收敛	定理10.3

10.5　参考资料

在线线性回归算法在 Strehl 和 Littman 的文章“Online linear regression and its application to model-based reinforcement learning”中有详细讨论. Bertsekas 和 Tsitsiklis 在 *Distributed and Parallel Computation:Numerical Methods* 中清晰讲解了分布式优化算法. 它解释了梯度投影算法和分布式实现，是一本很棒的参考书. LASSO 算法和其他许多方法及其应用在 Hastie 等的 *The Elements of Statistical Learning: Data Mining*，*Inference*、*and Prediction* , *2nd Edition* 中有详细阐述. 鞅理论则在其提出者 Doob 的书 *Stochastic Processes* 中有明确说明.

10.6　练习

1. 设 $\{Y_n, n \geqslant 1\}$ 为服从 $U[0,1]$ 的独立同分布随机变量，$\{Z_n, n \geqslant 0.1\}$ 为服从 $\mathcal{N}(0,1)$的独立同分布随机变量. 定义 $X_n = 1\{Y_n \geqslant a\} + Z_n, a$ 为常数. 设计一个算法，通过观

测点(X_n, Y_n)来学习 a 的值．采用以下的模型

$$X_n = g(Y_n - \theta),$$

其中，

$$g(u) = \frac{1}{1 + \exp\{-\lambda u\}} \tag{10.31}$$

且 $\lambda = 10$．注意，当 $u > 0$ 时，$g(u)$的分母接近 1，所以 $g(u) \approx 1$．同理，当 $u < 0$ 时，分母很大，所以 $g(u) \approx 0$．因此，$g(u) \approx 1\{u \geqslant 0\}$．函数 $g(\cdot)$叫作*逻辑函数*．在 Matlab 中运用随机梯度算法来估计 θ.

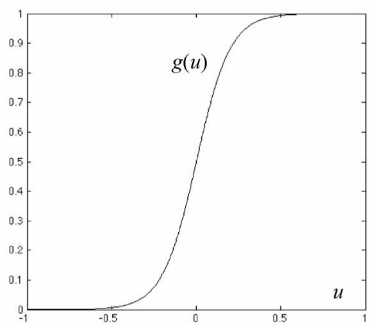


图 10-13　$\lambda = 10$ 时的逻辑函数（10.31）

2. 用下述值实现逐步回归算法

$$\Sigma_Z = \begin{bmatrix} 10 & 5 & 6 & 7 \\ 5 & 6 & 5 & 2 \\ 6 & 5 & 11 & 5 \\ 7 & 2 & 5 & 6 \end{bmatrix},$$

其中 $\boldsymbol{Z}' = (Y, X_1, X_2, X_3) = (Y, \boldsymbol{X}')$.

3. 用下式的信号实现压缩感知算法

$$s(t) = 3\sin(2\pi t) + 2\sin(3\pi t) + 4\sin(4\pi t), t \in [0,1].$$

假设在[0, 1]上均匀独立地选择采样时刻 t_k，并且正弦函数的频率为{0.1, 0.2, ⋯, 3}．为完全恢复原信号，至少需要几个采样点？

第11章

路线规划—A

应用：在不确定的延时情况下选择最优路线；控制马尔可夫链
主题：随机动态规划，马尔可夫决策问题

背景知识：第 1 章

本章介绍在面对不确定因素时如何作出决策，这些决策会影响每一步的代价和系统的状态．我们从一个简单的例子开始，在旅行时间不确定的情况下，选择一条最优路线；然后研究一个一般模型：控制马尔可夫链．



图 11-1　公路网络——如何选择路线

11.1　系统建模

给定一个有向连通图，其中包含有限个点，且图中的每一条边(i, j)都对应一个旅行时间 $T(i, j)$．旅行时间之间相互独立但分布已知．路线的起点是 s 节点，终点是 d 节点．我们的目标是选择一条最快捷的路线从 s 走到 d．下面会考虑几种解决方法．

为了让问题更加具体，我们来考虑图 11-2 所示的简单情况．

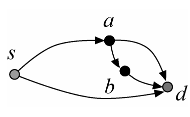


图 11-2　一个简单图

需要选择一条路线用最短的时间从 s 走到 d．在这个例子中，可能的路径为 sd、

sad 和 *sabd*．假设各边(i, j)对应的延时 $T(i, j)$如下：

$$T(s,a) =_D U[5,13], T(a,d) = 10, T(a,b) =_D U[2,10],$$
$$T(b,d) = 4, T(s,d) = 20.$$

也就是说,从 s 到 a 的延时均匀分布在[5, 13]区间内,从 a 到 d 的延时等于 10,等等.延时之间彼此独立（这是一个理想化的假设）.

11.2 方法 1：提前规划

这种方法不进行任何观测，只是在出发前事先计划旅行路线．此时认为每条边需要的时长为其平均旅行时间 $E(T(i, j)) = c(i, j)$，然后用最短路径算法找出一条路径．

如图 11-3 左上图所示，在我们的例子中，平均延时为 $c(s, a) = 9$、$c(a, d) = 10$，等等．

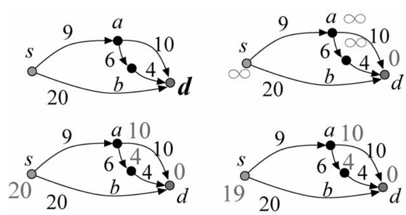


图 11-3 平均延时（左上图），以及贝尔曼–福特算法计算最小期望时间的步骤（浅色）

设 $V(i)$表示从节点 i 到目的地 d 的最小平均时间．贝尔曼–福特算法计算这些时间的方法如下．算法从 $V_0(d) = 0$ 和 $V_0(i) = \infty$ 开始，其中 $i \neq d$ ．设 $V_n(i)$ 为算法迭代计算 n 步后从 i 到 d 的最小平均时间估计值．之后，按照下式计算在第 n+1 次迭代之后的 $V_n(i)$ 值：

$$V_{n+1}(i) = \min_j\{c(i, j) + V_n(j)\}, n \geqslant 0, \tag{11.1}$$

其中 $V_n(i)$ 表示只经过至多 n 条边时从 i 到 d 的最小期望时间．如果经过至多 n 条边时没有可以达到目的地 d 的路径，则 $V_n(i)$ 的值为无穷大．9.2 节中的维特比算法也是基于该算法设计的．

这个算法的正确性基于：随机变量之和的期望等于它们的期望之和．例如，在至多经过 2 条边时，从 a 到 d 的最小平均旅行时间为 $V_2(a,d)$ ，相应的旅行时间为 $W_2(a,d)$ ．对于从 s 到 d 的平均时间最短的路线，如果只考虑至多经过 3 条边的路线，有两种可能：选择 *sd*，需要时间 $T(s, d)$；先选择 *sa*，然后选择至多经过 2 条边就能从 a 到 d 的最快路径，需要的时间为 $W_2(a,d)$ ．所以，至多经过 3 条边从 s 到 d 的最短期望时间 $V_3(s)$ 是 $E(T(s, d)) = c(s, d)$以及 $T(s, a)+W_2(a,d)$ 均值中的最小值，

$$V_3(s) = \min\{c(s,d), E(T(s,a) + W_2(a,d))\}$$
$$= \min\{c(s,d), c(s,a) + V_2(a,d)\}.$$

由于图中只有 N 个节点，任何一条通往 d 且不带环路的路径长度都不超过 N. 因此，V_n 可以在至多 N 步后收敛于 V. 极限值 $V(i)$表示的是从 i 到 d 的最短平均时间. V 满足下面的不动点等式：

$$V(i) = \min_j\{c(i,j) + V(j)\}, \forall j, V(d) = 0\,. \qquad (11.2)$$

这些方程就叫作动态规划方程. 式（11.1）是求解式（11.2）的一个算法.

11.3　方法 2：适应性算法

现在假设到达节点 i 时，能观测到从节点 i 出发的每一条路径需要的实际时间，不过仍然看不到更远的边所对应的时间. 这种情况下应该如何选择路线呢？如果每条边需要的时间是定值，那么这和方法 1 并没有什么区别. 然而，当它是随机变量时，从 i 出发的边所需要的实际旅行时间可能比期望值略大或略小. 显然我们应该用到这些信息.

下面给出一个计算最佳路径的系统流程. 设 $V(i)$是从节点 i 出发到达 d 的最小平均时间，$i \in \{s,a,b,d\}$. 可以看到 $V(b) = T(b, d) = 4$.

现在计算 $V(a)$. 已知从节点 a 出发所有边的延时，定义 $W(a)$为从 a 到 d 的最小期望时间，即

$$W(a) = \min\{T(a,b) + V(b), T(a,d)\}.$$

因此，$V(a) = E(W(a))$，即

$$V(a) = E(\min\{T(a,b) + V(b), T(a,d)\}). \qquad (11.3)$$

在这个例子中，我们有 $T(a,b) + V(b) =_D U[6,14]$. 因为 $T(a, d) = 10$，如果 $T(a,b) + V(b) < 10$（概率为 1/2），就选择路径 abd，这条路径的时间均匀分布在[6, 10]，均值为 8. 同时，如果 $T(a,b) + V(b) > 10$（概率也是 1/2），则选择旅行时间为 $T(a, d) = 10$ 的路线 ad. 因此，从 a 到 d 的最短旅行期望时间 $V(a)$取值为 8 和 10 的概率平均为 1/2. 均值为 $V(a) = 8(1/2)+10(1/2) = 9$.

类似地，

$$V(s) = E(\min\{T(s,a) + V(a), T(s,\mathrm{d})\}),$$

其中 $T(s,a) + V(a) =_D U[14,22]$，$T(s, d) = 20$. 如果 $T(s,a) + V(a) < 20$ 就选择从 s 到 a 的路径，这种情况出现的概率为(20−14)/(22−14) = 3/4. 这条路径所需的时间均匀分布在[14, 20]上，均值为 17. 如果 $T(s,a) + V(a) > 20$，那么选择直接的路线 sd，延时为

20．这种情况的概率是 1/4．因此，$V(s) = 17(3/4)+20(1/4) = 71/4 = 17.75$．

这里我们注意到，通过观察下一条边的延时并作出合理决策，我们将 s 到 d 的旅行时间期望值从 19 降低到了 17.5．这并不意外，因为信息越多越有帮助．同时，我们的决策取决于路径所需时间的观测值．例如，从节点 s 出发，如果$T(s,a)+V(a)>T(s,d)$，即$T(s,a)+9>20, T(s,a)>11$，就沿着边 sd 前进；否则选择 sa．

现在来考虑一般的情况．最关键的关系式如下：

$$V(i)=E(\min_j\{T(i,j)+V(j)\}), \forall i. \tag{11.4}$$

这个式子很好理解：从节点 i 出发，可以选择走到某个邻近的 j．这样的话，总的旅行时间就是从 i 到 j 的 $T(i, j)$加上之后从 j 到 d 的最小时间期望 $V(j)$．由于从 i 到 d 的路径必须通过另一个相邻的节点 j，所以从 i 到 d 的最短期望时间可以由上式给出．如前所述，这些等式的正确性是基于随机变量之和的期望等于它们的期望之和．

解决这些不动点方程的算法为

$$V_{n+1}(i)=E(\min_j\{T(i,j)+V_n(j)\}), n\geqslant 0. \tag{11.5}$$

对于所有的 i，都有$V_0(i)=0$．这里对$V_n(i)$的理解和提前规划类似：它表示的是经过最多 n 步时，从 i 到 d 所需的最短期望时间．当然，在路径上的每一步都可以观测到从当前结点出发各个边的延时．

方程组（11.4）是解决这个问题的随机动态规划方程．方程组（11.5）叫作迭代求值方程．

11.4 马尔可夫决策问题

更具一般性的路线决策问题是马尔可夫链的控制问题．在每一步都观察当前状态并作出下一步行为的决策，这一决策会决定转移概率以及下一步的代价．

具体来说，要在状态空间 $\mathcal{X}$ 上定义一个受控的马尔可夫链 $X(n)$，首先对于每一个 $x\in\mathcal{X}$ 都定义一个控制行为集合 $A(x)$，即在状态 x 下系统所能采用行为的集合．对于每一个状态 $x\in\mathcal{X}$ 以及行为 $a\in A(x)$，系统状态的转移概率为 $P(x,x';a)\geqslant 0$．这些概率满足 $\sum_{x'\in\mathcal{X}}P(x,x';a)=1$．同时，定义在状态 x 时采取行为 a 的代价为 $c(x, a)$．

在此设定下，序列 $X(n)$的定义如下：

$$\begin{aligned}&P[X(1)=x_1, X(2)=x_2,\cdots, X(n)=x_n\mid X(0)=x_0, a_0,\cdots,a_{n-1}]\\&=P(x_0,x_1;a_0)P(x_1,x_2;a_1)\times\cdots\times P(x_{n-1},x_n;a_{n-1}).\end{aligned}$$

我们要通过选择一系列的行为来最小化总代价的平均值：

$$E[\sum_{m=0}^{n}c(X(m), a(m))\mid X(0)=x]. \tag{11.6}$$

对于 $m = 0, \cdots, n$，行为 $a(m) \in A(X(m))$ 是由 $X(m)$、先前的状态 $X(0), \cdots, X(m-1)$和先前的行为 $a(0), \cdots, a(m-1)$决定的.

这个问题被称为马尔可夫决策问题.

解决这个问题的思路与之前的路径计划问题类似. 我们将这里的一个个状态当作路径中走过的节点. 设 $V_m(x)$ 表示将式（11.6）中的 n 替换成 m 时的代价最小值. 也就是说，$V_m(x)$ 是当初态为 $X(0) = x$ 时，下面 m+1 步平均代价的最小预期值. 这里的 $V_m(\cdot)$ 被称为价值函数.

其动态规划方程是

$$\begin{aligned} V_m(x) &= \min_{a \in A(x)} \{c(x,a) + E[V_{m-1}(x') \mid X(0) = x, a(0) = a]\} \\ &= \min_{a \in A(x)} \{c(x,a) + \sum_{x'} P(x,x';a)V_{m-1}(x')\}. \end{aligned} \tag{11.7}$$

设 $a = g_m(x)$ 是使得式（11.7）取值最小的 $a \in A(x)$ 的值，则 $a(m) \in g_m(X(m))$ 是使得式（11.6）取值最小的值.

当 $\mathcal{X}$ 和每一个 $A(x)$都是有限集时，一定存在行为 a 使式（11.7）取最小值. 该结论在更弱一些的假设下也成立.

示例

1. 猜测扑克牌

这里举一个简单的例子. 假设有一副已经完全洗均匀的扑克牌，共 52 张. 现在我们每次翻开一张牌. 在每次翻开下一张牌之前，你都有机会喊“停”. 如果下一张没有翻开的牌是 A，你就赢得 1 美元；如果不是，你就输了，游戏结束. 这个问题要探讨的是什么时候应该喊停.

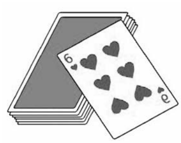


图 11-4　猜测下一张牌是不是 A

假设在剩余的 m 张牌当中还有 x 张 A. 那么如果喊停，赢的概率是 x/m. 如果不喊停，当下一张牌被翻起来之后，还剩下 x–1 张 A 的概率是 x/m，否则还剩下 x 张 A.

设 $V(x, m)$为在剩余的 m 张牌当中还有 x 张 A 的情况下，你获胜概率的最大期望值.

这里的动态规划方程是

$$V(m,x) = \max\{\frac{x}{m}, \frac{x}{m}V(m-1,x-1) + \frac{m-x}{m}V(m-1,x)\}.$$

有趣的是，可以证明以上方程的解是$V(m,x)=x/m$. 同时，当$x>0$时，取最大值操作的两项具有相同的值. 所以结论是，只要至少还有一张 A 在剩余的牌当中，随时喊停的获胜概率都相同.

2. 调度工作

假设你有两类工作要做. 第 i（i = 1, 2）类工作在完成之前，每单位等待时间的代价是c_i. 进行一项第 i 类工作的时候，无论你已经花了多少时间，在下一个时间单位完成工作的概率都是μ_i. 也就是说，每个 i 类工作所需的时长是几何分布的，参数为μ_i. 这里需要解决的问题是应当先做哪个工作才能使总的等待代价最小.

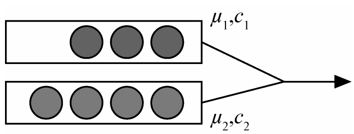


图 11-5　下一步该做什么工作

设$V(x_1, x_2)$表示有x_1个第 1 类工作、x_2个第 2 类工作的情况下，总等待代价的最小期望值. 与前面类似，可以得到以下的动态规划方程：

$$V(x_1,x_2)=x_1c_1+x_2c_2+\min\{\mu_1V((x_1-1)^+,x_2),\ \mu_2V(x_1,(x_2-1)^+)\}.$$

不妨假设$c_1\mu_1>c_2\mu_2$，可以直接证明动态规划方程的解如下：

$$V(x_1,x_2)=c_1\frac{x_1(x_1+1)}{2\mu_1}+c_2\frac{x_2(x_2+1)}{2\mu_2}+c_2\frac{x_1x_2}{\mu_1}.$$

而且，达到这个最小期望代价的方法是：总是先处理第 1 类工作. 这种策略被叫作 cμ 法则. 所以，尽管人们可能倾向于优先处理具有更多未完成任务的工作，但是这种选择却未必是最优的.

有一种互换的思路可以证明 cμ 法则的最优性. 假设你希望按照如下顺序完成工作：1221211. 也就是说，先完成一项 1 类工作，然后完成一项 2 类工作，之后再完成一项 2 类工作，以此类推. 我们按照如下方法改变策略：在完成第一项 2 类工作之后，转为处理原本在它之后的一项 1 类工作. 这么一来，就交换了第 1 类、第 2 类工作的顺序，也就改变了两类工作的等待时间（可以看到只有这两项工作受到影响）. 平均来看，第 1 类工作的等待时间降低了$1/\mu_2$，因为原先在它之前的第 2 类工作平均需要$1/\mu_2$的时间完成任务. 因此，第 1 类工作的等待代价降低了c_1/μ_2. 同理，第 2 类工作的等待代价平均增加了c_2/μ_1因此，总的平均代价降低了$c_1/\mu_2-c_2/\mu_1$.（因为$c_1\mu_1>c_2\mu_2$，所以这是一个正数）. 通过归纳，可以看到先做完所有第 1 类工作是最优的选择.

当然，能够通过简单推理选出最优策略的控制问题非常少．但是，仍然不要忘了这种可能性，它有时可以帮助我们简单地推导出结论．例如，假设工作排成如图 11-5 所示的队列，工作的到达过程服从独立的伯努利过程．也就是说，对于 $i = 1, 2$，第 i 类工作在每一个时间点到达的概率为 λ_i，彼此独立．用同样的互换策略可以证明：长期来看，cμ 法则能够使所有工作等待代价的平均期望值（并没有严格定义，不过读者可以猜出它的意思）最小．这很重要，因为这种情况下的动态规划方程并不能被显性地解出，而且严格证明这条法则的最优性非常复杂．

3. 雇用帮工

C.9

本节要考察的问题如下：假设需要处理的工作任务在随机的时间点到达，而你需要选择自己来做还是雇用帮工．直觉告诉我们，当待完成的任务量超过某一个阈值的时候，就应该需要一个帮工．

假设在每个时间点 $n = 0, 1, \cdots, a$ 有工作到达的概率是 $\lambda \in (0,1)$．如果自己处理，在单位时间内完成工作的概率为 $\mu \in (0,1)$，且该机率与过去的信息独立．如果雇用一个帮工，在单位时间内合作完成工作的概率为 $\alpha\mu \in (0,1)$，其中 $\alpha > 1$．设 $c(n) = \beta > 0$ 为雇用帮工时在时刻 n 的支出；而不雇用帮工的时候 $c(n) = 0$．我们需要最小化

$$E[\sum_{n=0}^{N}(X(n)+c(n))],$$

其中 $X(n)$表示在时刻 n 等待处理的工作总量．这个式子综合考虑了工作等待完成时的代价和雇用帮工的支出．如果一直雇用帮工，工作的等待代价就会最小；如果一直不雇帮工，雇用帮工的支出就会取得最小值．所以这个问题的核心是，找到雇用帮工的最佳时机，取得最优折中方案．

用 $X(n)$表示时刻 n 系统的状态．设

$$V_m(x) = \min E[\sum_{n=0}^{m}(X(n)+c(n)) \mid X(0)=x],$$

要在每一时刻根据系统的状态选择雇用帮工或不雇用，使得上式取最小值．随机动态规划方程为

$$\begin{aligned}V_m(x) = x + \min_{a\in\{0,1\}}\{&\beta 1\{a=1\} + (1-\lambda)(1-\mu(a))V_{m-1}(x)\\ &+ \lambda(1-\mu(a))V_{m-1}(\min\{x+1, K\})\\ &+ (1-\lambda)\mu(a)V_{m-1}(\max\{x-1, 0\})\\ &+ \lambda\mu(a)V_{m-1}(x)\}, n \geqslant 0,\end{aligned}$$

其中，定义 $\mu(0) = \mu$，$\mu(1) = \alpha\mu$，$V_{-1}(x) = 0$．同时，设积累的任务总量最大值为 K，即如果已经积累了 K 件工作任务，就不再接受新的任务．

我们用 Matlab 来解决这个问题．正如我们期望的，算法认为在时刻 n 如果

$X(n) > \gamma(N-n)$ 就应当雇一个帮工，其中 $\gamma(m)$ 是随着 m 递减的常数. 随着时间推移，m 增加，接受更多工作的代价增加，雇用帮工的倾向性也更大. 图 11-6 给出了 $\beta = 14$ 和 $\beta = 20$ 时 $\gamma(n)$ 的值. 图中 $\lambda = 0.5, \mu = 0.6, \alpha = 1.5, K = 20, N = 200$. 这里比较直观的是，帮工的工资越高，我们就会在积累了越大的工作量之后才决定雇用他们.

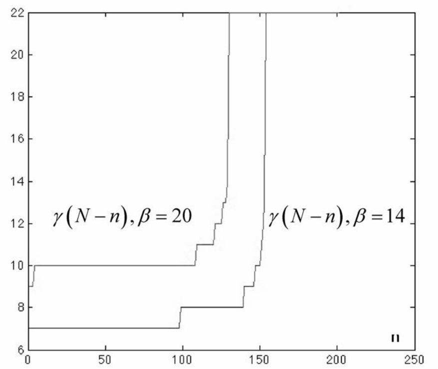


图 11-6　在时刻 n 如果积累的工作量超过 $\gamma(n)$ 就应当雇用帮工

4. 排哪个队

在超市采购以后，你会选择在哪个收银员的队伍中排队结账？一个很自然的做法是猜测哪个队的期望排队时间最短，然后排这个队. 每个人都会这样做，因为这种策略似乎会最小化每位顾客的排队时间. 但是你的一位朋友之前学过了这门课，他会告诉你，事情并不一定是这样的. 让我们一起来看看这个看似矛盾的说法吧.

假设有两个队伍，在每一时刻有顾客来排队的概率是 λ . 队列 i 的服务时间是关于参数 μ_i 的几何分布，其中 $i = 1, 2$.

假设当你到达时，第 i 个队伍中有 x_i 个顾客在等待（$i = 1, 2$）. 如果

$$\frac{x_1+1}{\mu_1} < \frac{x_2+1}{\mu_2},$$

你就应该加入队伍 1，因为这样会使得你等待服务的期望时间最短. 然而，如果希望使得两个队伍中所有顾客总的等待时间最短，就会发现公共最优策略和个人最优策略并不相同. 图 11-7 给出了 $\mu_2 < \mu_1$ 时的例子. 这张图说明在公共最优策略中有的顾客需要排在队伍 2，尽管这时他们的等待时间会大于采用个人最优策略时的等待时间.

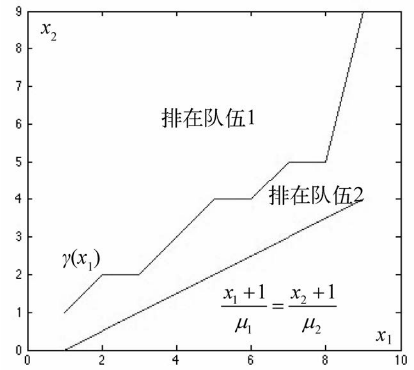


图 11-7　公共最优策略如折线所示，个人最优策略如直线所示

这个例子的目标可以被描述为最小化总的代价：

$$\sum_{n=0}^{N}\beta^n E(X_1(n)+X_2(n)).$$

在这个表达式中，$X_i(n)$ 表示时刻 n 在队伍 i 中排队的顾客数．每一个队伍的容量是 K．为了防止丢失太多顾客，我们要求如果只有一个队伍已经满员，则刚到达的顾客排到另一个未满的队伍．在总代价的表达式中，我们用折扣因子 $\beta\in(0,1)$ 来保证总代价有界．在图 11-7 中 $K=8, \lambda=0.3, \mu_1=0.3, \mu_2=0.2, N=100, \beta=0.95$．（事实上，图中的队伍长度是 x_1+1 和 x_2+2，因为 Matlab 不允许行标为 0.）

11.5　无限时域问题

求式（11.6）最小值时考虑的是有限的时长 n，即系统在时刻 n 停止．在这种情况下，我们研究了当系统处在状态 x、并还有 m 步要走时的最小代价 $V_m(x)$．因此，在任一时刻之后还要花费的代价以及在状态 x 的最佳行为均取决于剩余的时间．

当考虑的时长为无限时，问题反而变得简单了．这是因为剩余的时间在每个时刻都是相同的．这时候要使得总的代价有界，就需要对未来的代价进行折扣．因此，我们考虑的问题是最小化折扣后的代价：

$$E[\sum_{m=0}^{\infty}\beta^m c(X(m),a(m))\,|\,X(0)=x].\qquad(11.8)$$

在这个表达式中，$0<\beta<1$ 是折扣率．直观看来，如果 β 很小，那么未来的代价就不那么重要，人们也就不必考虑太远；而当 β 接近于 1 的时候，人们就需要对长期效应加以重视．

设 $V(x)$表示式（11.8）代价的最小值，即每一步可选行为都经过优化后的最小值．采用和前面类似的推导，可以证明

$$\begin{aligned} V(x) &= \min_{a\in A(x)}\{c(x,a)+\beta E[V(X(1))\mid X(0)=x,\,a(0)=a]\} \\ &= \min_{a\in A(x)}\{c(x,a)+\beta\sum_y P(x,y;a)V(y)\}. \end{aligned} \tag{11.9}$$

这里的式子与式（11.7）很像，不同之处在于折扣因子的存在，以及价值函数与时间无关．注意，这些式子也是不动点方程．求解这些方程的常见方法是考虑

$$V_{n+1}(x)=\min_{a\in A(x)}\{c(x,a)+\beta\sum_y P(x,y;a)V_n(y)\},n\geqslant 0, \tag{11.10}$$

其中$V_0(x)=0,\forall x$．这等价于：

$$V_n(x)=\min E[\sum_{m=0}^{n}\beta^m c(X(m),a(m))\mid X(0)=x].$$

可以证明式（11.10）的解$V_n(x)$满足$V_n(x)\to V(x),n\to\infty$，其中 $V(x)$为式（11.9）的解．

11.6 小结

- 动态规划方程
- 可控马尔可夫链
- 马尔可夫决策问题

重要方程与公式

马尔可夫决策问题	$P(x,y;a)$	11.4节
随机动态规划方程	$V_{m+1}(x)=E(\min_{a\in A(x)}\{c(x,a)+\sum_y P(x,y;a)V_m(y)\})$	式（11.7）

11.7 参考资料

Ross 的 *Introduction to Stochastic Dynamic Progamming* 一书是出色的随机动态规划入门书，“猜测扑克牌”的例子就是从中借鉴的．这本书简明易懂地阐述了核心概念，然后通过精心选择的例子讲解了定理的延伸和拓展．教材 *Dynamic Programming and Optimal Control* 全面地讲解了动态规划的算法．书中有很多例题，并且详细讨论了理论与实际的相关问题．

11.8 练习

1. 考虑一个离散的队列，有 1 名服务员为顾客提供服务. 每 1 个时间点有 1 位新顾客加入排队的概率为 $\lambda<1$. 服务员以 $\mu\in[0,1]$ 的速率为这个队伍工作，即每个时刻能完成 1 位顾客需求的概率是 μ. 由于精力有限，你希望服务员以尽可能小的速率提供服务，并保持队伍稳定. 因此，你希望服务员以 $\mu^*=\lambda$ 的速度工作. 不幸的是，你并不知道 λ 的具体值，能够看到的只是队伍的长度. 我们希望用基于随机化梯度的算法通过以下步骤得到 μ^* 的值.

(a) 使用梯度下降算法，调整 μ 使得函数 $V(p)=\dfrac{1}{2}(\lambda-\mu)^2$ 取值最小.

(b) 给定服务员在时间节点 n 的生产力 μ_n，计算 $E[Q(n+1)-Q(n)\,|\,Q(n)=q>0]$. $Q(n)$表示时刻 n 的队伍长度. 如果 $q=0$ 会怎样?

(c) 基于(a)和(b)的结论，使用随机梯度投影算法写出相应的 Matlab 代码. 请注意 $0\leqslant p\leqslant 1$.

提示：为了避免队列长度变为 0，从一个较大的队列长度开始计算.

2. 考虑有 3 个结点的路线图：起点是 s，终点是 d，中间结点是 r. 从 s 到 d 的直接路径耗时 20. 从 s 到 r 要 7 个单位时间. 从 r 到 d 有两条路径，其延时服从 8 到 20 之间的均匀分布，且相互独立.

(a) 如果希望在旅行之前作出规划，该选择哪一条路径从 s 走到 d?

(b) 如果在结点 r 能够确切知道从 r 到 d 需要的时间，应该选择哪条路?

3. 考虑一个离散时间的排队问题，其中新顾客的到达过程是一个参数为 λ 的伯努利过程. 有 1 名服务员以速率 μ 为这个队伍提供服务. 你可以考虑为队伍多分配 1 名服务员，这样能够使服务速率增加到 $\mu(1+a)$，$a>0$. 不过，增加 1 名服务员是有代价的. 设在时刻 n 的花费是 $X(n)+H(n)$，其中 $X(n)$是队列长度，$H(n)$是增加服务员的指示函数（增加一名服务员的时候等于 1，否则等于 0）. 你希望在有限时间情况下使代价最小化.

$$\sum_{n=0}^{N}E(X(n)+H(n))$$

(a) 写出动态规划方程.

(b) 取参数 $\alpha=0.5$ 且 $N=50$，用 Matlab 解决动态规划问题.

4. 对图 11-6 所示的路线图从节点 1 到 5 进行路线规划. 在图中各边上旅行所需的时间如下：$T(1,2)=2$、$T(1,3)\sim U[2,4]$、$T(2,4)=1$、$T(2,5)\sim U[4,6]$、$T(4,5)\sim U[3,5]$、$T(3,5)=4$. 其中 $X\sim U[\mathrm{a},\mathrm{b}]$表示随机变量 X 服从 a 到 b 的均匀分布.

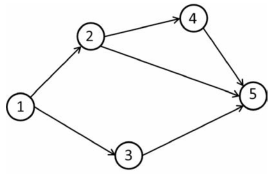


图 11-6　路线规划

(a) 如果想提前规划，应当选择哪一条路线？期望的旅行时间是多少？

(b) 现在假设在每一个结点，都可以知道下 2 步确切的旅行时间．就是说，在结点 1 可以知道除了 $T(4, 5)$以外所有路径的旅行时间．设 $V(i)$表示从结点 i 到 5 的最小期望时间，$1 \leqslant i \leqslant 5$．计算 $V(i)$，$1 \leqslant i \leqslant 5$．

5. 有一家“小盒子”工厂，主要做盒子生意．它在 k 年初存有 x_k 个盒子．根据合同，在 k 年末需要提供 d_k 个盒子．不过，d_k 的数值直到这一年结束才能确切知道．

在年初，他们可以计划生产 u_k 个盒子．通过使用廉价的劳动力，每个盒子的成本价是 A．在年末，“小盒子”工厂可以向“盒子大王”公司借 y_k 个盒子来完成合同的要求，代价是 $s(y_k)$．

满足合同要求后剩余的盒子可以留到下一年 $x_{k+1} = x_k + y_k - d_k$．可惜的是，这些多出来的盒子需要支付存储费 $r(x_{k+1})$．

现在你的任务是提供一个未来 20 年的盒子生产以及存储计划，目标是使得未来 20 年的总花销最小．设所有的花费在年末结算，没有通货膨胀．因为你在第 20 年就能拿到退休金了，所以不用担心 20 年以后“小盒子”工厂的花销．（不妨设开始时有 0 个盒子．）

(a) 从马尔可夫决策的角度，给这个问题建模．

(b) 写出动态规划方程．

(c) 用 Matlab 解上述方程，取如下参数：

- $r(x_k) = 5x_k$;
- $s(y_k) = 20y_k$;
- $A = 1$;
- $d_k =_D U\{1, 2, \cdots, 10\}$.

6. 甲和乙在比赛电子游戏，乙在时刻 0 从与甲相距 $T = 10$ 的位置出发，向甲靠近，速度是 1．例如，如果甲在平面的(0, 0)点，乙从(0, T)点出发向甲靠近，则经过 t 秒以后，乙的位置是(0，$T-t$)．甲选择了一个随机的时间点，均匀分布在[0, T]区间，此时他会向乙射击．如果甲先射击，乙就输了．设甲的射击一定能命中（这只是一个电子游戏的假设）.

(a) 乙需要决定在哪个时间点 t 向甲射击，使得自己获胜的概率最大．如果乙在与甲相距 x 的位置射击，他命中的概率为$1/(1+x)^2$．乙只有一次射击的机会．

(b) 求乙能够在决斗中获胜的最大概率？

(c) 现在假设乙有两颗子弹．请找到乙应当射击的两个时间点 t_1 和 t_2，使得他获胜的概率最大．同样，乙在与甲相距 x 的位置射击，命中的概率为$1/(1+x)^2$，两次射击相互独立．

7. 假设你在玩一个赌博游戏，获胜的概率 $p \in (0, 0.5)$，输的概率是 $1-p$．游戏开始时你有 16 枚金币，每次赌上 γ 枚金币，$\gamma \in \{1, 2, 4, 8, 16\}$．计算你在破产以前能够积攒到 256 枚金币的概率．初始资金为多大时，这个概率能够达到最大？

第12章

路线规划—B

主题：线性二次型高斯问题的控制，不完全观测

背景知识：第 7 章，第 11 章

12.1 线性二次型高斯问题

前面讨论的动态规划思想不仅可以应用在可控的马尔可夫链上，也可以用于其他可控制的系统中. 下面来考虑一个代价为平方函数且带有高斯白噪声的线性系统，即线性二次型高斯问题. 为了进行简化，我们只考虑标量的情况.

系统方程为

$$X(n+1)=aX(n)+U(n)+V(n), n\geqslant 0, \tag{12.1}$$

其中，$X(n)$表示系统状态，$U(n)$是一个控制量，$V(n)$是噪声. 设随机变量 $V(n)$为独立同分布的高斯变量 $N(0,\sigma^2)$.

这里需要解决的问题是，根据此前对系统状态的观测，在每一时刻 n 选择一个实数的控制值 $U(n)$，使得代价的期望值最小：

$$E[\sum_{n=0}^{N}(X(n)^2+\beta U(n)^2)\,|\,X(0)=x]. \tag{12.2}$$

控制的目标是使得系统状态的值接近于 0，并限制为此付出的代价.

问题的核心是寻找两方面的折中：一方面是系统状态值很大时的代价，另一方面是控制系统状态接近于 0 的代价. 我们先通过一个简化的问题来获得对求解的直观认识. 考虑最小化：

$$(ax+u)^2+\beta u^2.$$

这个简单问题没有噪声影响，而且只进行一次控制. 为了得到这个（关于 u 的）表达式的最小值，我们将这个式子关于 u 求偏导，并令偏导数为 0，得到

$$2(ax+u)+2\beta u=0,$$

所以

$$u=-\frac{a}{1+\beta}x.$$

可见，使得花销最小的控制量与系统状态成线性关系．当系统状态远远偏离 0 时，需要的控制量很大．下面的定理表明，同样的结论也适用于式（12.2）.

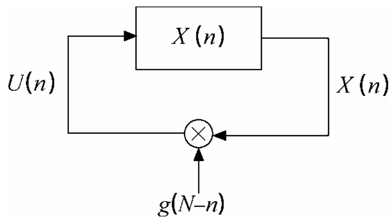


图 12-1　最佳控制与系统状态成线性关系

定理 12.1　最佳线性二次型高斯控制

对于式（12.1）所示的系统，使得式（12.2）取最小值的控制量 $U(n)$为

$$U(n)=g(N-n)X(n).$$

其中 $d(-1)=0$，

$$g(m)=-\frac{ad(m-1)}{\beta+d(m-1)}, m\geqslant 0; \quad (12.3)$$

$$d(m)=1+\frac{a^2\beta d(m-1)}{\beta+d(m-1)}, m\geqslant 0. \quad (12.4)$$

最佳控制与系统状态成线性关系，且系数取决于剩余的时间．这些系数可以在时刻 0 计算出来，而且不受噪声的影响．可以假设 $V(n)=0$，然后计算所有的控制量.

证明 □

设 $V_m(x)$ 为式（12.2）的最小值，将其中的 N 换成 m．随机动态规划方程为

$$V_m(x)=\min_u\{x^2+\beta u^2+E(V_{m-1}(ax+u+V))\}, m\geqslant 0, \quad (12.5)$$

其中 $V=N(0,\sigma^2)$，取 $V_{-1}(x)=0$.

可以验证这些方程的解为

$$V_m(x) = c(m) + d(m)x^2,$$

其中 $c(m)$和 $d(m)$为常量，$d(m)$满足式（12.4）.

也就是说，

$$\min_u \{x^2 + \beta u^2 + E[c(m-1) + d(m-1)(ax+u+V)^2]\} = c(m) + d(m)x^2. \quad (12.6)$$

其中 $d(m)$由式（12.4）给出，且最优取值为 $u = g(m)x$，其中 $g(m)$由式（12.3）给出.

这里的证明是一个简单的代数计算，请读者自行证明. ♣

令$N \to \infty$

如果（12.2）中的 N 变得非常大会怎么样？观察式（12.4），我们发现如果$|a| < 1$，那么$d(m) \to d$ 且 $m \to \infty$，其中 d 是以下不动点方程的解：

$$d = f(d) := 1 + \frac{a^2 \beta d}{\beta + d}.$$

为了说明原因，请注意

$$f'(d) = \frac{a^2 \beta^2}{(\beta + d)^2},$$

所以，对于所有的 $d \geqslant 0$，都有 $0 < f'(d) < a^2$ 且 $f(d) > 0$．因此，$f(d)$ 是一个压缩映射：

$$|f(d_1) - f(d_2)| \leqslant \alpha |d_1 - d_2|, \forall d_1, d_2 \geqslant 0,$$

其中$\alpha \in (0,1)$（这里$\alpha = a^2$）. 现在取$d_1 = d$ 且 $d_2 = d(m)$，我们得到：

$$|d - d(m+1)| \leqslant \alpha |d - d(m)|, \forall m \geqslant 0.$$

因此

$$|d - d(m)| \leqslant \alpha^m |d - d(0)|,$$

这表明$d(m) \to d$，如前文所述. 所以，式（12.3）表明$m \to \infty$时$g(m) \to g$，其中

$$g = -\frac{ad}{\beta + d}.$$

所以，当剩余的时间还很多的时候，最佳的控制值接近于 $U(N-m) = gX(N-m)$. 这表明这个控制策略可能在 N 趋于无穷大时最小化式（12.2）所示的花销.

解决这个问题的正式做法是考虑下式定义的长期平均花销：

$$\lim_{N \to \infty} \frac{1}{N} E[\sum_{n=0}^{N} (X(n)^2 + \beta U(n)^2) | X(0) = x].$$

这个表达式是单位时间内的平均花销. 可以证明，如果$|a| < 1$，那么控制值取 $U(n) =$

$gX(n)$，一定可以使平均花销最小，其中 g 的定义如前所述.

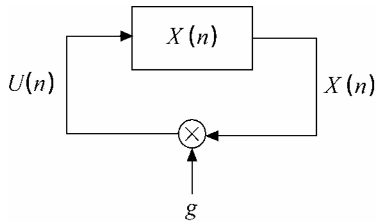


图 12-2　达到平均花销的最佳控制值

12.2　有噪声观测时的线性二次型高斯问题

在上一小节，我们控制了一个具有高斯白噪声但系统状态可观测的线性系统．下面考虑系统状态观测带有噪声的情况.

系统可以描述为

$$X(n+1)=aX(n)+U(n)+V(n), n\geqslant 0; \quad (12.7)$$

$$Y(n)=X(n)+W(n). \quad (12.8)$$

其中随机变量 $W(n)$为独立同分布的$\mathcal{N}(0,w^2)$变量且与 $V(n)$独立.

这里要解决的问题是，在每一时刻 n，基于 $Y^n:=\{Y(0),\cdots,Y(n)\}$的值找到合适的控制值 $U(n)$，使得总花销式（12.2）的期望值最小.

下面的定理给出了这个问题的解.

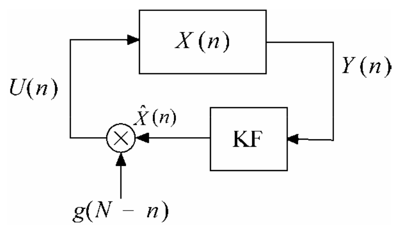


图 12-3　最佳控制值与系统状态的期望成线性关系

定理 12.2　有噪声观测时的最优线性二次型高斯控制问题

问题的解为

$$U(n)=g(N-n)\hat{X}(n).$$

其中

$$\hat{X}(n)=E[X(n)\,|\,Y(0),\cdots,Y(n),U(0),\cdots,U(n-1)]$$

可以用卡尔曼滤波器以及式（12.3）~（12.4）定义的常数 $g(m)$计算出来.

因此，这里的控制值与 $X(n)$能够准确观测到时的计算值一样，只是把 $X(n)$替换成了 $\hat{X}(n)$. 这个性质被称为**确定性等价**. □

证明

定理中 $g(n)$的取值与噪声 $V(n)$无关这一点也给了我们一些理解定理内容的提示：给定 Y^n，系统状态 $X(n)$是方差为 v^2 的随机变量 $\mathcal{N}(\hat{X}(n),v^2)$. 于是，可以将噪声观测看作增加了系统状态的方差，就像是 $V(n)$的方差增加了一样.

下面阐述这个结论的大致证明思路，详细的推导过程省略. 设在时刻 $N-m+1$ 给定 Y^{N-m+1} 情况下未来花销的最小期望值为

$$c(m-1)+d(m-1)\hat{X}(N-m+1)^2,$$

则在 $N-m$ 时刻，给定 Y^{N-m} 以及 $U(N-m)=u$，未来花销的期望值就是下式的期望值：

$$X(N-m)^2+\beta u^2+c(m-1)+d(m-1)\hat{X}(N-m+1)^2.$$

现在有

$$X(N-m)=\hat{X}(N-m)+\eta.$$

其中 η 是与 Y^{N-m} 独立的高斯随机变量. 同时，与卡尔曼滤波器的情况一样，

$$\hat{X}(N-m+1)=a\hat{X}(N-m)+u$$
$$+K(N-m+1)\{Y(N-m+1)-E[Y(N-m+1)\,|\,Y^{N-m}]\}.$$

此外，根据联合高斯分布的条件期望，$Y(N-m+1)-E[Y(N-m+1)\,|\,Y^{N-m}]$是一个期望为零且与 Y^{N-m} 独立的高斯随机变量. 因此，

$$\hat{X}(N-m+1)=a\hat{X}(N-m)+u+Z.$$

其中 Z 是一个均值为 0 的高斯随机变量.

于是，在时刻 $N-m+1$，未来花销的期望值为下式的期望：

$$(\hat{X}(N-m)+\eta)^2+\beta u^2+c(m-1)+d(m-1)(a\hat{X}(N-m)+Z)^2,$$

即

$$\hat{X}(N-m)^2+\beta u^2+c(m-1)+d(m-1)(a\hat{X}(N-m)+u+Z)^2.$$

这与式（12.6）一致，只是把 x 换成了 $\hat{X}(N-m)$，把 V 换成了 Z. 最后，由于 V 的方差大小并不影响 $c(m)$和 $d(m)$的值，这就证明了我们的结论. ♣

令 $N\to\infty$

与 $X(n)$能够精确观测的情况一样，可以证明，如果$|a|<1$，则控制量

$$U(n) = g\hat{X}(n)$$

能够使单位时间内的平均花销最小. 同时, 这里的卡尔曼滤波器变成了稳态的滤波器, 具有如下形式:

$$\hat{X}(n+1) = a\hat{X}(n) + u + K[Y(n+1) - a\hat{X}(n) - U(n)].$$

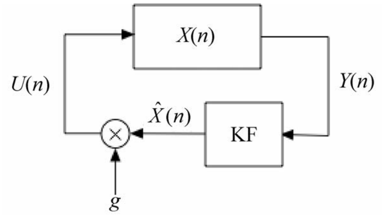


图 12-4　根据噪声观测，使平均花销最小的最佳控制值. 这里滤波器为稳态卡尔曼滤波器

12.3　部分可观测的马尔可夫决策问题

前面章节讨论过可控的马尔可夫链，其中每一时刻的控制都基于已知的系统状态. 在这一节，我们研究马尔可夫链的状态无法准确观测的问题，即可控的隐马尔科夫链. 这一类问题被称为部分可观测的马尔可夫决策问题.

我们先来看一些具体的例子，为读者提供这些问题的基本思路. 稍后讨论一般性的情况.

示例

找钥匙

假设你找不到钥匙了. 不过你知道钥匙要么在 A 包中，要么在 B 包中. 钥匙在 A 包的概率为 p. 不幸的是，你的包里堆满了杂乱的东西. 如果花一个单位时间（例如 10 秒钟）在 A 包中寻找，而且钥匙确实在 A 包内，找到的概率是 α. 类似地，如果钥匙确实在 B 包中，花一个单位时间在其中寻找，能够找到的概率为 β. 在每一个时刻，都可以选择下一步寻找哪个包. 你的目标是使得找到钥匙所花时间的期望值最小.

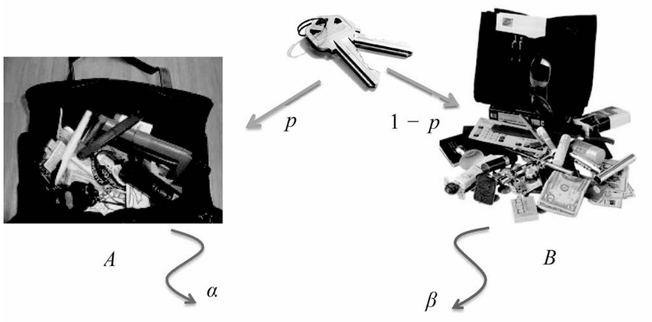


图 12-5　钥匙在哪里

这个系统的状态是钥匙的位置，即 A 或 B．但你无法观测到这个状态．核心的思路是考虑在已知时刻 n 以前所有观测量的条件下，钥匙在 A 包中的条件概率 p_n．这里的 p_n 是一个可控马尔可夫链．不幸的是，p_n 的取值集合为[0,1]，这是一个无限不可数的集合．不过先别泄气．

假设在时刻 n，钥匙在 A 包中的条件概率为 p_n，你在 A 包中找了 1 个单位时间，并没有找到钥匙．那么 p_{n+1} 是多少呢？我们说

$$p_{n+1} = \frac{p_n(1-\alpha)}{p_n(1-\alpha)+(1-p_n)} =: f(A, p_n).$$

式中的分子是钥匙在 A 包中但没有被找到的概率，分母是没有找到钥匙的概率（钥匙可能在 A 包中也可能在 B 包中）．当然，如果找到了钥匙，问题就结束了．

类似地，假设你在 B 包中找了一番，没有找到．那么

$$p_{n+1} = \frac{p_n}{p_n+(1-p_n)(1-\beta)} =: f(B, p_n).$$

因此，你的行动在控制着 p_n．设 $V(p)$表示找到钥匙所花时间的最小期望值，其中 p 表示钥匙在 A 包中的概率．则相应的动态规划方程为

$$V(p) = 1 + \min\{(1-p\alpha)V(f(A,p)), (1-(1-p)\beta)V(f(B,p))\}. \tag{12.9}$$

式中的常数 1 表示第一步花费的时间．取最小值的第一项表示你在 A 包中寻找钥匙的情况：找不到钥匙的概率为$1-p\alpha$，此时需要等待最小期望时间 $V(f(A,p))$ 才能找到钥匙，因为此时钥匙在 A 包中的概率变为 $f(A,p)$．另一项对应于先在 B 包中寻找钥匙的情况．

这些方程看起来简直无从下手，不过很容易通过 Matlab 来求解．将[0,1]区间离散成 K 个子区间，在循环中更新 $f(A,p)$ 和 $f(B,p)$ 的值．于是，通过更新得到的值是一个有限长度的向量 $\boldsymbol{V}$={$V(1/K)$,$V(2/K)$,⋯,$V(1)$}．在这种离散情况下，方程组（12.9）可以写成

$$\boldsymbol{V}=\phi(\boldsymbol{V}).$$

其中 $\phi(\cdot)$ 表示式（12.9）右侧的式子. 这些是不动点方程. 为了求解，取初始值 $\boldsymbol{V}_0=\boldsymbol{0}$，按下式迭代：

$$\boldsymbol{V}_{t+1}=\phi(\boldsymbol{V}_t), t\geqslant 0.$$

如果运气比较好的话，可以通过数学方法来证明动态规划方程的解，即这种算法能够收敛到 $\boldsymbol{V}$. 解的情况如图 12-6 所示，其中 α 和 β 有不同的取值. 图中将最佳的行动决策绘成 p 的函数. 代码如图 12-7 所示. 在离散化的方法中，我们在[0,1]区间内取了 K=1000 个点并迭代了 100 次.

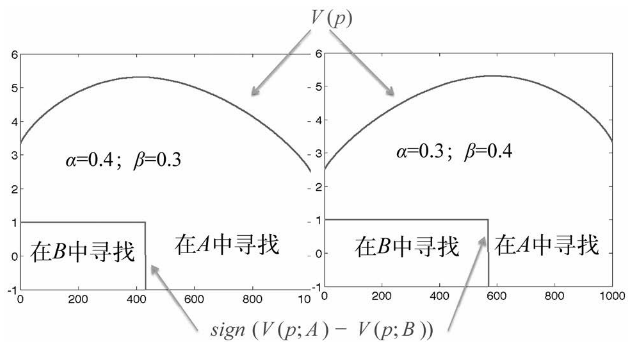


图 12-6　式（12.9）的数值解（另见彩插）

```
% 寻找钥匙
a = 1;
b = 0.3;
N = 1000; %
T = 100; %
ac = 1 - a;
bc = 1 - b;
V = 100*ones(1, N); % will store the value function
K = V;
for p = 1:N
    K(p) = p;
end
VA = V;
VB = V;
for n = 1:T
    for p = 1:N
        VA(p) = 1 + (1 - a*p/N)*V(min(round(0.5 + ac*p/(1 - a*p/N)), N));
        VB(p) = 1 + (1 - b*(1 - p/N))*V(min(round(0.5 + p/(1 - b*(1 - p/N))), N)),
        V(p) = min(VA(p), VB(p));
    end
end
plot(K,V,K,sign(VA - VB))
```

图 12-7　解式（12.9）的 Matlab 代码

12.4　小结

- 能够观测系统状态的线性二次型高斯控制问题

- ❑ 有噪声观测的线性二次型高斯控制问题
- ❑ 部分可观测的马尔可夫决策问题

重要方程与公式

线性二次型高斯问题	规范化	式（12.1）~（12.2）
线性二次型高斯问题求解	$U_n = g_{N-n}X_n$	定理12.1
有噪声观测	$Y_n = X_n + W_n$	式（12.8）
有噪声观测的问题求解	$U_n = g_{N-n}\hat{X}_n$	定理12.2
部分可观测的马尔可夫决策问题	将 X_n 替换成 $P\left[X_n = x \middle| Y^n\right]$	12.3节

12.5 参考资料

Bertsekas 的 *Dynamic Programming and Optional Control*，Kumar 和 Varaiya 的 *Stochastic Systems:Estimation*, *Identification and Adaptive Control* 以及 Goodwin 和 Sin 的 *Adaptive Filtering Prediction and Control* 都探讨了线性二次型高斯控制问题．前两本书也讨论了部分可观测的马尔可夫决策问题.

12.6 练习

1. 考虑如下系统

$$X(n+1) = 0.8X(n) = U(n) + V(n),\ \ n \geqslant 0.$$

其中 $X(0)=0$，随机变量 $V(n)$独立同分布于 $N(0, 0.2)$．$U(n)$是控制量.

(a) 取 $U(n)$=0， $n \geqslant 0$ ，为这个系统进行仿真.

(b) 取 N=100，用定理 12.1 给出的控制值为系统仿真.

(c) 用具有常数增益的控制量 $g = \lim_{n\to\infty} g(n)$ 为系统仿真.

2. 考虑如下系统

$$X(n+1) = 0.8X(n) = U(n) + V(n),\ \ n \geqslant 0$$
$$Y(n) = X(n) + W(n), n \geqslant 0.$$

其中 $X(0)=0$，随机变量 $V(n)$与 $W(n)$相互独立，$V(n) =_D \mathcal{N}(0,0.2)$ 且 $W(n) =_D \mathcal{N}(0,\sigma^2)$.

(a) 取 $\sigma^2 = 0.1$ 和 $\sigma^2 = 0.4$ ，用定理 12.2 给出的控制量为系统仿真.

(b) 取 $\sigma^2 = 0.1$ 和 $\sigma^2 = 0.4$ ，用增益取极限值时的控制量以及卡尔曼滤波器为系统仿真.

(c) 比较上述时变控制与极限控制下的系统.

3. 有两枚硬币，一枚均匀，另一枚正面向上的概率是 0.6. 只通过外观是无法分辨出两枚硬币的. 在第 $n \geqslant 1$ 步，你可以选择抛掷一枚硬币，使得正面向上次数的期望值最大.

(a) 用部分可观测的马尔科夫决策问题为系统建模.

(b) 仿照“找钥匙”问题的解决步骤，将系统的状态离散化，写出离散化的随机动态规划方程.

(c) 在 Matlab 中应用该随机动态规划方程，为得到的系统仿真.

第13章

视野拓展和补充

视野拓展：推断问题
补充：充分统计量，无限马尔可夫链，泊松过程，连续时间马尔可夫链，容量，边界，鞅，强大数定理

我们已经探索了由实际应用引出的一些主题. 现在，让我们把这些概念结合起来，查漏补缺.

13.1 推断问题

我们探讨过的一个关键概念是推断. 推断问题有如下的规范化描述：有一对随机变量(X, Y)，我们希望根据观察到的 Y 来猜测 X.

$$Y \overset{?}{\rightarrow} X$$

图 13-1　推断问题就是从 Y 的值猜测 X 的值

因此，需要找到一个函数 $g(\cdot)$使得 $\hat{X} := g(Y)$在某种精确的定义上接近 X. 下面是一些例子.

- X 是某人的体重，Y 是其身高.
- $X = 1$ 代表一栋着火的房子，没有着火则 $X = 0$，而 Y 是探测器量出的 CO 浓度.
- $X \in \{0, 1\}^N$ 是发射机发送的比特串，$Y \in \Re^{[0,T]}$ 是接收机接到的信号.
- Y 是一位女性的基因组，如果她会患某种乳腺癌则 $X = 1$，反之 $X = 0$.
- Y 是描述一部电影和一个人的特性向量，X 是此人给电影的评分.
- Y 是某人的脸部照片，如果是男性则 $X = 1$，反之 $X = 0$.
- X 是一句话，Y 是麦克风接收到的信号.

我们在第 5 章和第 7 章中已经解释了一些不同类型的问题.

- 已知分布：知道(X, Y)的联合分布.
- 离线：观测一系列样本点(X, Y)的值.
- 在线：连续观测样本(X, Y)的值.
- 最大似然估计：不假定 X 的先验分布，只假定 Y 在给定 X 时的条件分布；目标是找到 X 的值使得观测值 Y 最可能发生.
- 最大后验估计：知道 X 的先验概率以及 Y 在给定 X 下的条件分布；目标是找到 X 的值，使得观测值为 Y 的可能性最大.
- 假设检验：不假设 $X\in\{0, 1\}$的分布，只假设 Y 对 X 的条件分布；目标是使正确决策 $X=1$ 的概率最大，并且保证当 $X=0$ 时，判定 $X=1$ 的概率不大于 β.
- 最小均方误差：给定 X 和 Y 的联合分布密度，想要找到函数 $g(Y)$使 $E((X-g(Y))^2)$最小.
- 最小线性均方估计：给定 X 和 Y 的联合分布密度，想要找到线性函数 $a+bY$，使得 $E((X-a-bY)^2)$最小.

13.2　充分统计量

推断问题中的一个重要概念是充分统计量．到目前为止，我们还没有讨论过这一概念．现在让我们来看看它的定义.

定义 13.1　充分统计量

如果对于变量 X，$h(Y)$满足

$$f_{Y|X}[y\,|\,x]=f(h(y),x)g(y),$$

另一个等价的描述为

$$f_{Y|h(Y),X}[y\,|\,s,x]=f_{Y|h(Y)}[y\,|\,s],$$

则称 $h(Y)$是一个 X 的**充分统计量**.　◊

我们将这一等价性的证明留给读者.

在讨论定义之前，首先挖掘一下定义的含义．如果我们有先验分布 $f_X(x)$并且想要计算 $MAP[X|Y=y]$，则

$$\begin{aligned}MAP[X\,|\,Y=y]&=\arg\max_x f_X(x)f_{Y|X}[y\,|\,x]\\&=\arg\max_x f_X(x)f(h(y),x)g(y)\\&=\arg\max_x f_X(x)f(h(y),x).\end{aligned}$$

所以，使 MAP 最大的是 $h(y)$的一个函数．也就是，对某一函数 $g(\cdot)$，有

$$MAP[X|Y] = g(h(Y)).$$

简单来说，用于计算 $MAP[X|Y]$的 Y 的信息全部包含在 $h(Y)$中．

同样，$MLE[X|Y]$也是一个 $h(Y)$的函数．同时，我们还注意到：

$$f_{X|Y}[x \mid y] = \frac{f_X(x) f_{Y|X}[y \mid x]}{f_Y(y)} = \frac{f_X(x) f(h(y), x) g(y)}{f_Y(y)}.$$

由此可得：

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_X(x) f(h(y), x) g(y)\, dx = g(y) \int_{-\infty}^{\infty} f_X(x) f(h(y), x)\, dx \\ &= g(y)\phi(h(y)), \end{aligned}$$

其中，

$$\phi(h(y)) = \int_{-\infty}^{\infty} f_X(x) f(h(y), x) dx.$$

因此，

$$f_{X|Y}[x \mid y] = \frac{f_X(x) f(h(y), x)}{\phi(h(Y))}.$$

所以，在给定 Y 时 X 的条件密度只取决于 $h(Y)$．于是，

$$E[X \mid Y] = \psi(h(Y)).$$

现在再考虑当 $X \in \{0, 1\}$时的假设检验问题．注意，

$$L(y) = \frac{f_{Y|X}[y \mid 1]}{f_{Y|X}[y \mid 0]} = \frac{f(h(y), 1) g(y)}{f(h(y), 0) g(y)} = \psi(h(y)).$$

所以，似然率同样只取决于 $h(y)$，由此推出假设检验问题的解也是一个 $h(Y)$的函数．

解释

充分统计量的定义十分抽象．直观的理解是，如果 $h(Y)$对 X 是充分的，那么 Y 是关于 $h(Y)$和一个独立于 X、Y 的随机变量 Z 的函数．也就是，

$$Y = g(h(Y), Z). \tag{13.1}$$

比如，假设 $Y = (Y_1, \cdots, Y_n)$，其中 Y_m 是依 $x \in [0, 1]$独立同分布的伯努利随机变量．令 $h(Y) = Y_1 + \cdots + Y_n$．那么，我们可以把 Y 想像成在$(Y_1, \cdots, Y_n)$中随机选取 $h(Y)$个等于 1 的分量而构造出的向量．这一选择根据某个独立随机变量 Z 完成．在这种情况下，可以看到 Y 中关于 X 的信息也都包含在 $h(Y)$中．

为了理解这一解释与定义之间的等价性，首先假设式（13.1）成立．那么，

$$P[Y \approx y \big| X = x] = P[h(Y) \approx h(y) \big| X = x] P(g(h(y), Z) \approx y)$$
$$= f(h(y), x) g(y),$$

所以，$h(Y)$对于 X 是充分的．反过来，如果 $h(Y)$对于 X 是充分的，则能找到某个变量 Z 使得 $g(h(y), Z)$等于 $f_{Y|h(Y)}[y|h(y)]$.

13.3 无限马尔可夫链

我们学习了在有限空间 $\mathcal{X} = \{1, 2, \cdots, N\}$上定义的马尔可夫链，接下来探究无限的情形，其中 $\mathcal{X} = \{0, 1, \cdots\}$.

首先给定初始分布 $\pi = \{\pi(x), x \in \mathcal{X}\}$，其中 $\pi(x) \geqslant 0$ 且 $\sum_{x\in\mathcal{X}} \pi(x) = 1$．另外，我们还给定了一个非负数集合$\{P(x, y), x, y \in \mathcal{X}\}$使得

$$\sum_{y\in\chi} P(x, y) = 1, \ \forall x \in \mathcal{X}.$$

若

$$P(X(0) = x_0, X(1) = x_1, \cdots, X(n) = x_n)$$
$$= \pi(x_0) P(x_0, x_1) \times \cdots \times P(x_{n-1}, x_n)$$

对于 $n \geqslant 0$ 和 $\mathcal{X}$ 中的所有 $x_0, \cdots, x_n$ 都成立，则序列$\{X(n), n \geqslant 0\}$是一个初始分布为 π 的马尔可夫链且其概率转移矩阵为 P.

与有限马尔可夫链的情形相同，定义“不可约”和“非周期”的定义．在之前的讨论中，如果一个有限马尔可夫链是不可约的，那么它会无限次访问其所有状态，且在每个状态上花费的长期时间比例为正数.

但是当马尔可夫链是无限时，上述情况可能不会发生．为了理解这一点，考虑下面的例子（见图 13-2）．我们有 $\pi(0) = 1$ 和 $P(i, i+1) = p (i \geqslant 1)$，且

$$P(i+1, i) = 1 - p =: q = P(0, 0), \forall i.$$

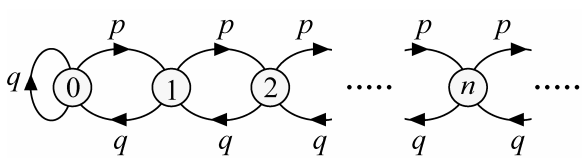


图 13-2 一个无限马尔可夫链

假设 $p \in (0, 1)$，不难看到马尔可夫链是不可约的．若 $p > 0.5$，可以清楚地看到，当 $n \to \infty$时，$X(n) \to \infty$．为了说明情况的确如此，令 $Z(n)$为独立同分布随机变量，$P(Z(n) = 1) = p$ 且 $P(Z(n) = -1) = q$．则有

$$X(n) = \max\{X(n-1) + Z(n), 0\},$$

所以

$$X(n) \geqslant X(0) + Z(1) + \cdots + Z(n-1), n \geqslant 0.$$

而且，

$$\frac{X(n)}{n} \geqslant \frac{X(0) + Z(1) + \cdots + Z(n-1)}{n} \to E(Z(n)) > 0,$$

这里的收敛由强大数定理得到．这意味着如前所述，$X(n) \to \infty$．

所以，$X(n)$最终会比给定的任何 N 都大，并且会越来越大．这说明 $X(n)$访问每个状态的次数是有限的．这时称这些状态是瞬态的．

如果一个状态不是瞬态的，则称它是递归的．在这种情况下，如果连续访问一个状态所间隔的平均时间是有限的，那么这个状态叫作正递归；否则叫作零递归．

下面是与定理 1.2 相对应的结果．

定理 13.2　无限马尔可夫链大定理

考虑一个无限马尔可夫链．

(a) 如果该马尔可夫链是不可约的，其状态可能是全瞬态、全正递归或者全零递归．在这些情况下，分别称马尔可夫链是瞬态的、正递归的和零递归的．

(b) 如果该马尔可夫链是正递归的，它会有唯一的稳态分布 π，其中 $\pi(i)$ 是 $X(n)$等于 i 的长期时间比例．

(c) 如果该马尔可夫链是正递归的、非周期的，那么 $X(n)$的分布 π_n 收敛于 π．

(d) 如果该马尔可夫链不是正递归的，则其没有不变分布且在任何状态停留的时间比例趋于 0． □

对于图 13-2 的马尔可夫链，$p = 0.5$ 时是零递归的，$p < 0.5$ 时是正递归的．在后一种情况中，稳态分布是

$$\pi(i) = (1-\rho)\rho^i, i \geqslant 0，\text{其中 } \rho := \frac{p}{q}.$$

李雅普诺夫–福斯特准则

这是一个非常有用的判断正递归的充分条件．

定理 13.3　李雅普诺夫–福斯特准则

令 $X(n)$为一个定义在无限状态空间 $\mathcal{X}$ 上的不可约马尔可夫链．假设存在

非负函数 V：$\mathcal{X}\rightarrow[0,\infty)$使得

$$E[V(X(n+1))-V(X(n))\big|X(n)=x]\leqslant-\alpha+\beta 1\{x\in A\},$$

其中 A 为一有限集，$\alpha>0$ 且 $\beta>0$；则该马尔可夫链是正递归的．函数 V 叫作马尔可夫链的李雅普诺夫函数．　□

这种情况意味着，当 $X(n)$在某有限集 A 外时，李雅普诺夫函数平均减量至少为 α. 这一结果表明马尔可夫链是正递归的，原因如下：由于李雅普诺夫函数是非负的，不可能一直递减，所以它在有限集 A 内停留的时间比例必然为正数．根据大定理可知，这意味着它是正递归的.

13.4　泊松过程

泊松过程在应用概率论中是一个重要模型．它能很好地近似数据包到达路由器的过程、电话呼叫的到来、新 TCP 连接的产生以及收银台处顾客的出现.

13.4.1　定义

我们从泊松过程的定义开始讨论（见图 13-3）.

定义 13.4　泊松过程

令 $\lambda>0$ 且$\{S_1, S_2, \cdots\}$为 Exp(λ)的独立同分布随机变量．又令 $T_n=S_1+\cdots+S_n$，$n\geqslant 1$．定义

$$N_t=\max\left\{n\geqslant 1\middle|T_n\leqslant t\right\},t\geqslant 0,$$

其中，当 $t<T_1$ 时，$N_t=0$．则 $\boldsymbol{N}:=\{N_t, t\geqslant 0\}$是一个参数为 λ 的泊松过程．T_n 是 $\boldsymbol{N}$ 的第 n 个跳跃的时间．　◊

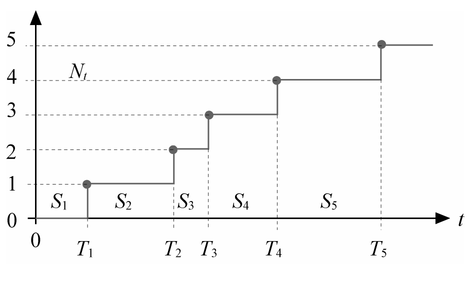


图 13-3　泊松过程：跳转之间的时间 S_n 是参数为 λ 的独立同分布指数随机变量

13.4.2 独立自增量

在探讨泊松过程的性质之前，我们先来回顾指数分布的两个性质.

定理 13.5 指数分布的性质

令 τ 为指数分布参数为 $\lambda>0$ 的变量，则，

$$F_\tau(t)=P(\tau\leqslant t)=1-\exp\{-\lambda t\}, t\geqslant 0.$$

对于 $t\geqslant 0$，τ 的分布密度函数为 $f_\tau(t)=\lambda\exp\{-\lambda t\}$. 且 $E(\tau)=\lambda^{-1}$，$\mathrm{var}(\tau)=\lambda^{-2}$. 那么，我们有

$$P[\tau>t+s|\tau>s]=P(\tau>t).$$

这是指数分布的**无记忆性质**.

同时，

$$P[\tau\leqslant t+\varepsilon|\tau>t]=\lambda\varepsilon+o(\varepsilon).$$ □

证明

$$\begin{aligned}P[\tau>t+s|\tau>s]&=\frac{P(\tau>t+s)}{P(\tau>s)}\\&=\frac{\exp\{-\lambda(t+s)\}}{\exp\{-\lambda s\}}=\exp\{-\lambda t\}\\&=P(\tau>t).\end{aligned}$$ ♣

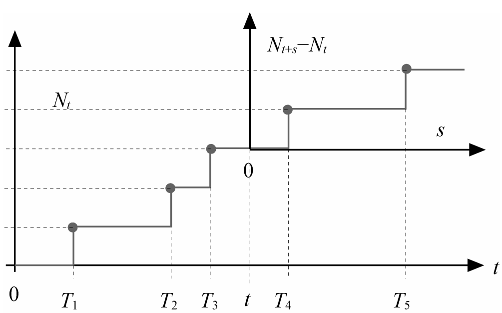


图 13-4 给定该过程到时刻 t 前的情况，时刻 t 以后的跳跃过程构成了一个泊松过程

该性质的解释是，如果一个灯泡的使用寿命是指数分布的，那么一个旧的灯泡和一个新的灯泡完全一样（只要它还亮着）.

下面根据这个性质用严格的推导证明泊松过程也是无记忆的.

定理 13.6　泊松过程的无记忆性

令 $\boldsymbol{N} := \{N_t, t \geqslant 0\}$ 是一参数为 λ 的泊松过程. 固定 $t > 0$，给定 $\{N_s, s \leqslant t\}$，过程 $\{N_{s+t} - N_t, s \geqslant 0\}$ 是参数为 λ 的泊松过程.

因此，该过程有稳定独立的增量. 也就是，对于任意 $0 \leqslant t_1 < t_2 < \cdots$，泊松过程的增量 $\{N_{t_{n+1}} - N_{t_n}, n \geqslant 1\}$ 是独立的，且分布 $N_{t_{n+1}} - N_{t_n}$ 只取决于 $t_{n+1} - t_n$.

□

证明

图 13-4 阐明了这一结果. 给定 $\{N_s, s \leqslant t\}$，第一次跳跃 $\{N_{s+t} - N_t, s \geqslant 0\}$ 的时间是 $\mathrm{Exp}(\lambda)$，可从指数分布的无记忆性得出. 接下来的跳跃次数是独立同分布随机变量，服从参数为 λ 的指数分布. 这就证明了该定理. ♣

13.4.3　跳跃次数

我们有下面的结论.

定理 13.7　跳跃次数是泊松分布

$\boldsymbol{N} := \{N_t, t \geqslant 0\}$ 是参数为 λ 的泊松过程. 那么 N_t 服从均值为 λt 的泊松分布.

□

证明

有很多证明这一结论的方法，其中的标准方法如下. 注意到，

$$P(N_{t+\varepsilon} = n) = P(N_t = n)(1 - \lambda\varepsilon) + P(N_t = n-1)\lambda\varepsilon + o(\varepsilon).$$

因此，

$$\frac{\mathrm{d}}{\mathrm{d}t} P(N_t = n) = \lambda P(N_t = n-1) - \lambda P(N_t = n).$$

从而得到

$$\frac{\mathrm{d}}{\mathrm{d}t} P(N_t = 0) = -\lambda P(N_t = 0).$$

由初始条件 $P(N_0 = 0) = 1$，得到对于 $t \geqslant 0$，$P(N_t = 0) = \exp\{-\lambda t\}$. 现在，假设

$$P(N_t = n) = g(n,t)\exp\{-\lambda t\}, n \geqslant 0.$$

那么上面的微分方程变为

$$\frac{\mathrm{d}}{\mathrm{d}t}[g(n,t)\exp\{-t\}] = \lambda[g(n-1,t) - g(n,t)]\exp\{-\lambda t\},$$

也就是，

$$\frac{\mathrm{d}}{\mathrm{d}t}g(n,t)=\lambda g(n-1,t).$$

由递归可得 $g(n,t)=\dfrac{(\lambda t)^n}{n!}$.

另一种证明用到了跳跃的密度．如前述定义，令 T_n 为该过程的第 n 次跳跃时间，并令 $S_n = T_n - T_{n-1}$．那么，

$$\begin{aligned}
&P(T_1\in(t_1,t_1+\mathrm{d}t_1),\cdots,T_n\in(t_n,t_n+\mathrm{d}t_n),T_{n+1}>t)\\
&=P(S_1\in(t_1,t_1+\mathrm{d}t_1),\cdots,S_n\in(t_n-t_{n-1},t_n-t_{n-1}+\mathrm{d}t_n),S_{n+1}>t-t_n)\\
&=\lambda\exp\{-\lambda t_1\}\mathrm{d}t_1\lambda\exp\{-\lambda(t_2-t_1)\}\mathrm{d}t_2\cdots\exp\{-\lambda(t-t_n)\}\\
&=\lambda^n\mathrm{d}t_1\cdots\mathrm{d}t_n\exp\{-\lambda t\}.
\end{aligned}$$

为得到这个表达式，我们用到了 S_n 是服从 Exp(λ)的独立同分布随机变量．上式告诉我们，如果已知[0, t]之间有 n 次跳跃，则它们出现在区间任一位置的可能性相等．而且，

$$P(N_t=n)=\int_S\lambda^n\mathrm{d}t_1\cdots\mathrm{d}t_n\exp\{-\lambda t\},$$

其中，$S=\{t_1, \cdots, t_n|0<t_1<\cdots<t_n<t\}$．现在，注意到 S 是$[0, t]^n$ 中对应次数 t_i 按特定顺序排列的子集．一共有 n!个顺序，且由对称性可知，每一个顺次对应于$[0, t]^n$ 中相同大小的子集．所以，S 的大小为 $t^n/n!$．我们可以得出

$$P(N_t=n)=\frac{t^n}{n!}\lambda^n\exp\{-\lambda t\},$$

从而证明了这一结果． ♣

13.5 连续时间马尔可夫链

前面已经学习了离散时间马尔可夫链$\{X(n), n=0, 1, 2, \cdots\}$．在这一节，我们要对连续时间的情况的主要思想进行概述．

设 $\mathcal{X}$ 为一可数的或有限的集合．我们定义过程$\{X_t, t\geqslant 0\}$如下：给定一个定义在 $\mathcal{X}$ 上的概率分布 π 和速率矩阵 $Q=\{Q(i,j), i,j\in\mathcal{X}\}$．

由定义，Q 满足

$$Q(i,j)\geqslant 0,\forall i\neq j \text{ 且 } \sum_j Q(i,j)=0,\forall x.$$

定义 13.8 连续时间马尔可夫链

如果过程$\{X_t, t\geqslant 0\}$满足 $P(X_0=i)=\pi(i)$以及

$$P[X_{t+\varepsilon}=j \mid X_t=i, X_u, u<t]=1\{i=j\}+\varepsilon Q(i,j)+o(\varepsilon).$$

则称其为初始分布为 π、速率矩阵 Q 的连续时间马尔可夫链. ◊

这个定义意味着该过程从状态 i 跳到 $j\neq i$ 的概率在 $\varepsilon \ll 1$ 个时间单位内为 $Q(i, j)\varepsilon$. 所以，$Q(i,j)$是每个时间单位内从 i 跳到 j 的概率. 注意到，这些表达式对 j 求和得 1.

下面是一种构建这种过程的方法. 假设 $X_t = i$. 选取一个指数分布的随机时间 τ，其速率 $q(i):=-Q(i,i)$. 在时间 $t+\tau$ 处,该过程跳跃到状态 y 的概率对 $j\neq i$ 为 $\Gamma(i,j)=Q(i,j)/q(i)$.

因此，若 $X_t = i$，概率 $X_{t+\varepsilon} = j$ 就是过程在$(t, t+\varepsilon)$中跳跃的概率，等于 $q(i)\varepsilon$，乘以接下来跳到 j 的概率，也就是 $\Gamma(i,j)$. 所以，

$$P[X_{t+\varepsilon}=j \mid X_t=i]=q(i)\varepsilon\frac{Q(i,j)}{q(i)}=Q(i,j)\varepsilon,$$

误差为 $o(\varepsilon)$. 因此，这一构造能得出正确的转换.

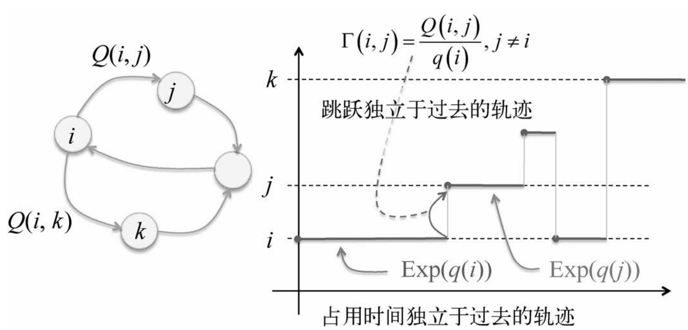


图 13-5 构建连续时间马尔可夫链

下面是与定理 13.2 相对应的结论. 这里“不可约”的定义与前面一样. 同时，在连续时间情况下，没有周期的概念.

定理 13.9 连续时间马尔可夫链大定理

考虑一个连续时间马尔可夫链.

(a) 如果该马尔可夫链是不可约的，其状态可能是全瞬态、全正递归或者全零递归. 分别称这个马尔可夫链是瞬态的、正递归的和零递归的.

(b) 如果该马尔可夫链是正递归的，它会有唯一的稳态分布 π，其中 $\pi(i)$ 是 X_t 等于 i 的长期时间比例. 而且，马尔可夫链 X_t 在状态 i 的概率 $\pi_t(i)$收敛于 $\pi(i)$.

(c) 如果该马尔可夫链不是正递归的，则其没有稳态分布且在任何状态

停留的时间比例都趋于 0. □

13.6 二元对称信道的容量

考虑一个出错概率为 $p \in (0, 0.5)$的二元对称信道．因为发射机发送的每一个比特都有可能出错，所以信道不可能完全可靠．不论发送端传送什么，接收机都无法确定它接收到的信息是正确的．

尽管如此，我们还是可以达到一个很小的出错率．比如，不妨设 $p = 0.1$，并通过重复 N 次来传送一个比特，其中 $N \gg 1$．接收机收到这 N 个比特后采用多数译码．也就是说，如果它收到的 0 多于 1，就判断发送机发送的是 0，反之则为 1．通过选取很大的 N，出错率可以变得任意小．由于这种方法每经过 N 步才传送一个比特，所以该信道的速率是 $1/N$．这似乎意味着，为了实现很小的出错率，我们必须牺牲速率.

幸运的是，这种悲观的结论是错误的．在 20 世纪 40 年代，克劳德・香农揭示了信道可以用比 $C(p)$小的任意速率传送信息，其中（见图 13-6），

$$C(p) = 1 - H(p), H(p) = -p\log_2 p - (1-p)\log_2(1-p), \tag{13.2}$$

且对于任意 $\varepsilon > 0$，传输的出错概率小于 ε .

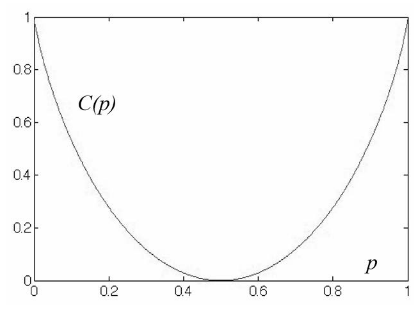


图 13-6 二元对称信道的容量 $C(p)$的出错概率为 p

比如，$C(0.1) \approx 0.53$．现在选定一个比 $C(0.1)$小的速率，如 $R = 0.5$．选择任意 $\varepsilon > 0$，设 $\varepsilon = 10^{-8}$．那么我们能用速率 $R = 0.5$ 来传送比特，且每个比特的出错概率小于 10^{-8}．选择 $\varepsilon = 10^{-12}$ 时也成立：可以用同样的速率 R 来传送，且出错概率小于 10^{-12}．实际使用的传输机制取决于 ε，当 ε 更小的时候规则会变得更复杂；但是速率 R 不取决于 ε．这个结论很惊人！不用说，它让之前忙于设计各种基于直觉的传送方案的工程师们困惑不已．



图 13-7　克劳德·香农（1916—2001）

香农最重要的见解是，*长序列是很典型的*. 在如马尔可夫链或者独立同分布随机变量的序列中，存在一个统计上的规律性，且该规律性在长序列中尤为明显. 比如，多次抛掷一枚有偏差的硬币，假设 $P(\text{正面朝上}) = 0.1$，观察到的序列很可能有 10%的正面. 而许多其他的序列几乎不可能出现，你几乎看不到. 所以，实际可能出现的长序列相对来说很少. 在这个例子中，当 $P(\text{正面朝上}) = 0.1$ 时，尽管抛掷 N 次硬币有 $M = 2^N$ 个可能出现的序列，但是只有大概 $\sqrt{M}$ 个是典型的. 而且，由对称性可知，所有这些典型序列以等概率出现. 正是这个原因使得二元对称信道的错误模式相对较少. 不妨设每 N 次传送仅有 A 种可能的错误模式. 那么，每个长度为 N 的比特串会对应于 A 个可能接收到的输出串：每个典型错误对应一个序列. 因此，可以为发送机选择 B 个长度为 N 的输入串使得其中每一个输入串对应的 A 个输出串都不同. 然而，有人也许会担心，由于想让输出串的集合不同，选择这 B 个输入串会很困难.

然而，香农注意到，这一目标可以通过完全随机地选取输入串来实现. 香农的方案如下. 选取一个较大的 N，然后通过抛一枚均匀的硬币 N 次来随机选取 B 个 N 比特的字串. 记这些输入串为 $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_\text{B}$，称为*码字*. 令 S_1 为对应于输入 $\boldsymbol{X}_1$ 的 A 个典型输出的集合. 令 $\boldsymbol{Y}_j$ 为对应于输入 $\boldsymbol{X}_j$ 的输出. 由信道的对称性可知，$\boldsymbol{Y}_j$ 是一个均匀硬币的抛掷序列. 所以，每个 $\boldsymbol{Y}_j$ 为都以相同的概率等于 2^N 个可能输出字串中的某一个. 其中 $\boldsymbol{Y}_j$ 落在 S_1 中的概率为 $A/2^N$. 实际上，

$$P(\boldsymbol{Y}_2 \in S_1 \text{或} \boldsymbol{Y}_3 \in S_1 \cdots \text{或} \boldsymbol{Y}_B \in S_1) \leqslant B \times A2^{-N}.$$

在上述不等式中，用到了事件集合的概率不大于其概率之和的性质. 我们下面会解释 $A = 2^{NH(p)}$. 所以，如果选择 $B = 2^{NR}$，可以看到上式的 P 小于等于

$$2^{NR} \times 2^{NH(p)} \times 2^{-N}.$$

同时，如果

$$R+H(p) < 1，\text{即 } R < C(p) := 1-H(p),$$

那么当 N 增大的时候，这一数值趋近于 0.

所以，如果不选择过多的码字，接收机出错的概率可以忽略. 同时注意到 $B = 2^{NR}$ 等同于在 N 步内传送 NR 个不同的比特，因此传送速率为 R.

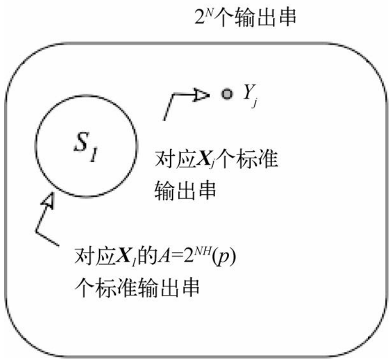


图 13-8　由于码字为随机选取的，一个码字产生另一个码字典型输出的概率为 $A2^{-N}$

接收机要如何识别发射机发送的比特串呢？思想是把 B 个输入串（码字）的列表给接收机．当接收机接收到一个串，它从列表中寻找与其最相近的串．这个找到的字串非常可能就是发送机传送的字串．

我们还需要解释为什么 $A=2^{NH(p)}$．幸运的是，这个结果可以简单地由强大数定理推出．令 $\boldsymbol{X}:=\{X(n), n=1, \cdots, N\}$为独立同分布随机变量，其中 $P(X(n)=1)=p$ 且 $P(X(n)=0)=1-p$．对于一个给定的序列 $\boldsymbol{x}=(x(1), \cdots, x(N))\in\{0,1\}^N$，令

$$\psi(\boldsymbol{x}):=\frac{1}{N}\log_2(P(\boldsymbol{X}=\boldsymbol{x})). \tag{13.3}$$

注意到，通过令 $|\boldsymbol{x}|:=\sum_{n=1}^{N}x(n)$，有

$$\begin{aligned}\psi(\boldsymbol{x})&=\frac{1}{N}\log_2(p^{|\boldsymbol{x}|}(1-p)^{N-|\boldsymbol{x}|})\\&=\frac{|\boldsymbol{x}|}{N}\log_2(p)+\frac{N-|\boldsymbol{x}|}{N}\log_2(1-p).\end{aligned}$$

所以，N 比特的随机串 $\boldsymbol{X}$ 使得

$$\psi(\boldsymbol{X})=\frac{|\boldsymbol{X}|}{N}\log_2(p)+\frac{N-|\boldsymbol{X}|}{N}\log_2(1-p).$$

由强大数定律可知，当 $N\to\infty$时，$|\boldsymbol{X}|/N\to p$．因此，对于 $N\gg 1$，

$$\psi(\boldsymbol{X})\approx p\log_2(p)+(1-p)\log_2(1-p):=-H(p).$$

这一计算告诉我们，对于 $\boldsymbol{X}$ 可能取值的任意序列 $\boldsymbol{x}$，$\psi(\boldsymbol{x})$的值都相同．由式（13.3）可得，这意味着所有可能的序列 $\boldsymbol{x}$ 几乎有相同的概率 $2^{-NH(p)}$．我们概括这一现象如下：存在 $2^{NH(p)}$个等概率的典型序列．所以，$A=2^{NH(p)}$．

回顾一下高斯信道的最大似然估计检测规则，满足下述条件时，信道变为二元对称信道

$$p = p(\sigma^2) := P(\mathcal{N}(0, \sigma^2) > 0.5).$$

于是，可以计算出容量 $C(p(\sigma^2))$，它是噪声标准偏移量 σ 的函数．图 13-9 显示出了结果．

香农在信道容量和可达传输速率上得到的结果对于通信系统的设计具有深远的影响．突然间，工程师们有了目标，知道自己的系统离可实现的最佳速率还差多远．香农的编码机制尽管并不实用，却为信道编码设计提供了重要的见解．香农的信息论展现了深邃的理论如何在工程领域引起一场革命．

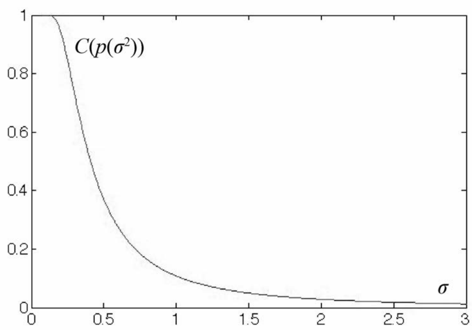


图 13-9　在附加噪声 $\mathcal{N}(0, \sigma^2)$下的二元对称信道的容量．检测器采用最大似然估计

香农另一个重要的工作是关于随机事物的编码．比如，编码一本 500 页的书要用多少比特？在这里，典型性再次发挥了重要的作用．例如，我们知道编码一个抛掷 N 次偏差率为 P(正面朝上) = p 硬币的字符串只需要 $NH(p)$比特，因为这是典型序列所需要的比特数．在这里，$H(p)$叫作抛硬币的熵．类似地，如果$\{X(n), n \geqslant 1\}$是一个不可约、有限且非周期性的马尔可夫链，并具有稳态分布 π 和状态转移矩阵 $P(i, j)$．那么，我们大致需要 $NH(p)$比特来对$\{X(1), \cdots, X(N)\}$进行编码，其中

$$H(P) = -\sum_i \pi(i) \sum_j P(i, j) \log_2 P(i, j)$$

叫作马尔可夫链的熵率．一个名为 Liv-Zempel 的实用压缩机制基本上可以实现这一极限．这也是大多数文件压缩算法（比如，ZIP）的基础．

香农还将信道容量和源编码两个想法结合起来，这个结果就是克劳德・香农的分离定理．我们用下面的例子来解释其信源信道编码的结果．考虑由马尔可夫链生成的信号 $X(n)$通过一个二元对称信道传输．那么最快的速率能达到多少？答案是 $C(p)/H(p)$．这个结果可以这样理解：描述每一个符号要用 $H(p)$比特，而二元对称信道可以在每个时间单位传送 $C(p)$比特．而且，为了达到这一速率，我们先对信源进行编码，然后单独选择二元对称信道的编码，最后一起运用．所以，信道编码不依赖于信源编码，反之亦然．

13.7 概率界

下面讲解如何运用切比雪夫不等式、切尔诺夫不等式和高斯估计来得出概率的估计值．这些方法同时也告诉我们相应事件发生的可能性．这些方法的重要之处在于，它们可以应用于分析非常复杂的情况．



图 13-10　赫尔曼・切尔诺夫（1923—　）

定理 13.10　马尔可夫、切尔诺夫和詹森不等式

令 X 为随机变量，那么我们有

(a) **马尔可夫不等式**：设 $f(\cdot)$ 为非减的正函数，则

$$P(X \geqslant a) \leqslant \frac{E(f(X))}{f(a)}. \tag{13.4}$$

(b) **切尔诺夫不等式**：对于所有 $\theta > 0$，有

$$P(X \geqslant a) \leqslant E(\exp\{\theta(X-a)\}). \tag{13.5}$$

(c) **詹森不等式**：设 $f(\cdot)$ 为凸函数，则

$$f(E(X)) \leqslant E(f(X)). \tag{13.6}$$

□



图 13-11　乔汉・詹森（1859—1925）

这一定理的证明很简单，如下.

证明

(a) 因为 $f(\cdot)$是非减的正函数，所以

$$1\{X \geqslant a\} \leqslant \frac{f(X)}{f(a)},$$

两边同时取期望即可得到式（13.4）.

(b) 不等式（13.5）是马尔可夫不等式（13.4）的特殊情况，取 $f(X) = \exp\{\theta X\}$，$\theta > 0$.

(c) 设 $f(\cdot)$为凸函数，即它在其所有切线之上，

$$f(X) \geqslant f(E(X)) + f'(E(X))(X - E(X)),$$

如图 13-12 所示．现在对两边同时取期望即得不等式(13.6).

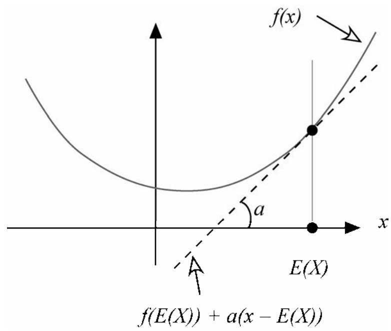


图 13-12　凸函数 $f(\cdot)$在其所有切线之上，所以也在 $E(X)$处的切线之上，从而得到詹森不等式

概率界在多址复用中的应用 ♣

回顾前面介绍的多址复用问题．存在 N 位相互独立的用户，每位用户传输的概率为 p．因此，在线用户的数目 Z 服从 $B(N, p)$分布．我们想要找到值 m，使得 $P(Z \geqslant m) = 5\%$.

第一种估计 m 的方法运用切比雪夫不等式（2.2），即

$$P(|v - E(v)| > \varepsilon) \leqslant \frac{\text{var}(v)}{\varepsilon^2}.$$

若 $Z = B(N, p)$，则 $E(Z) = Np$ 且 $\text{var}(Z) = Np(1-p)$[①]．由于 $v = B(100, 0.2)$，我们得到 $E(v) = 20$ 且 $\text{var}(v) = 16$．由切比雪夫不等式可得

$$P(v - 20 > \varepsilon) \leqslant \frac{16}{\varepsilon^2}.$$

① 参见附录 A.

我们有理由认为 v 的分布在其均值附近几乎是对称的，如图 3-4 所示．所以，我们得到

$$P(v-20>\varepsilon)\leqslant\frac{8}{\varepsilon^2},$$

现在要选择 $m=20+\varepsilon$，使得 $P(v>m)\leqslant 5\%$．这意味着我们要选择 ε，使得 $8/\varepsilon^2=5\%$．所以，$\varepsilon=13$，$m=33$．由切比雪夫不等式，可以认为不会有超过 33 名用户同时在线．那么在设计网络的时候，可以选择 C 使得 $C/33$ 是一个能让用户满意的速率．

第二种方法运用切尔诺夫不等式（13.5），也就是

$$P(v\geqslant Na)\leqslant E(\exp\{\theta(v-Na)\}),\forall\theta>0.$$

为了计算右边的大小，我们注意到，如果 $Z=B(N,p)$，那么可以将 Z 写作 $Z=X(1)+\cdots+X(N)$，其中 $X(n)$是独立同分布随机变量且 $P(X(n)=1)=p$，$P(X(n)=0)=1-p$．因此，

$$\begin{aligned}E(\exp\{\theta Z\})&=E(\exp\{\theta X(1)+\cdots+\theta X(N)\})\\&=E(\exp\{\theta X(1)\}\times\cdots\times\exp\{\theta X(N)\}).\end{aligned}$$

我们还注意到，由于 $X(n)$相互独立，随机变量 $\exp\{\theta X(n)\}$[①]也相互独立．而且，独立随机变量之积的期望等于它们期望的乘积（见附录 A）．因此，

$$\begin{aligned}E(\exp\{\theta Z\})&=E(\exp\{\theta X(1)\})\times\cdots\times E(\exp\{\theta X(N)\})\\&=E(\exp\{\theta X(1)\})^N=\exp\{N\Lambda(\theta)\}\end{aligned}$$

在上式中，我们定义

$$\Lambda(\theta)=\log(E(\exp\{\theta X(1)\})).$$

因此，由切尔诺夫不等式可得

$$\begin{aligned}P(Z\geqslant Na)&\leqslant\exp\{N\Lambda(\theta)\}\exp\{-\theta Na\}\\&=\exp\{N(\Lambda(\theta)-\theta a)\}\end{aligned}$$

由于不等式对于任意 $\theta>0$ 都成立，我们将右边看作 θ 的函数并寻找其最小值．定义

$$\Lambda^*(a)=\max_{\theta>0}\{\theta a-\Lambda(\theta)\}.$$

那么，

$$P(Z\geqslant Na)\leqslant\exp\{-N\Lambda^*(a)\}.\tag{13.7}$$

图 13-13 展示了 $p=0.2$ 时该函数的状态．

① 实际上，独立随机变量的函数也是独立的，见附录 A．

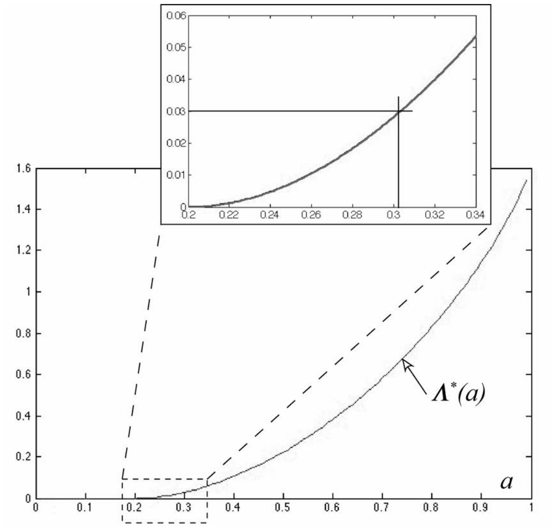


图 13-13　许多在线用户概率界的对数的 N 倍，即 Λ

现在来计算 $\Lambda(\theta)$ 和 $\Lambda^*(a)$．我们得到：

$$E(\exp\{\theta X(1)\}) = 1 - p + p\mathrm{e}^{\theta},$$

所以，

$$\Lambda(\theta) = \log(1 - p + p\mathrm{e}^{\theta})$$

且

$$\Lambda^*(a) = \max_{\theta>0}\{\theta a - \log(1 - p + p\mathrm{e}^{\theta})\}.$$

对其求导并对 θ 的偏导数取零，我们发现

$$a = \frac{1}{1 - p + p\mathrm{e}^{\theta}}(p\mathrm{e}^{\theta}),$$

当 $a > p$ 时，

$$\mathrm{e}^{\theta} = \frac{a(1-p)}{(1-a)p}.$$

最后将其带回 $\Lambda^*(a)$，得到

$$\Lambda^*(a) = -a\log\left(\frac{a}{p}\right) - (1-a)\log\left(\frac{1-a}{1-p}\right), \forall a > p.$$

回到我们的例子，希望找到值 $m = Na$ 使得

$$P(v \geqslant Na) \approx 0.05.$$

运用式（13.7），需要找到 Na 使得

$$\exp\{-N\Lambda^*(a)\} \approx 0.05 = \exp\{\log(0.05)\},$$

也就是，

$$\Lambda^*(a) = -\frac{\log(0.05)}{N} \approx 0.03.$$

参照图 13-13，可以得到 $a = 0.30$．这对应于 $m = 30$．所以，切尔诺夫的估计告诉我们 $P(v > 30) \approx 5\%$．我们可以依据这一结果来规划网络的容量．

还能看到，上面的计算说明切尔诺夫界也可以写作

$$P(Z \geqslant Na) \leqslant \frac{P(B(N,p) = Na)}{P(B(N,a) = Na)}.$$

13.8 鞅

鞅代表一个人玩公平投机游戏时的财富值．在这种游戏中，盈利的期望值永远是 0．一个简单的例子是以 0 为步长均值的随机游走．鞅是很好的噪声模型，同时也能很好地对依期望值衰减的过程进行建模（比如，股票市场）．鞅理论最早由杜步提出 [Doob，1953]．

鞅有一个重要的性质，就是将强大数定理进行了扩展：一个期望值有界的鞅几乎必然收敛．这一结论可以用于说明随机过程的浮动会消失，最终收敛于其平均值．随机梯度算法的收敛性以及用微分方程近似随机过程的方法均是由该定理得来．

13.8.1 定义

当某人玩投机游戏时，令 X_n 为时间 $n \geqslant 0$ 时该玩家的财富．如果该游戏满足

$$E[X_{n+1} \mid X^n] = X_n, \forall n \geqslant 0. \tag{13.8}$$

则认为游戏是公平的．

在这个表达式中，$X^n := \{X_m, m \leqslant n\}$．因此，在一个公平的游戏中，没有人能指望自己的财富增加．由满足这一性质的随机变量构成的序列 $\{X_n, n \geqslant 0\}$ 为鞅．

上述定义可以扩展到当玩家还拥有其他额外信息但是仍然无法增加自己财富的情况．比如，不妨设额外信息是另一个随机变量 Y_n 的值．根据这个思路，可以得到下面的定义．

定义 13.11　鞅、上鞅、下鞅

若

$$E[X_{n+1} \mid X^n, Y^n] = X_n, \forall n \geqslant 0, \tag{13.9}$$

则随机变量序列$\{X_n, n\geqslant 0\}$是对于$\{X_n, Y_n, n\geqslant 0\}$的一个鞅. 其中$X^n = \{X_m, m\leqslant n\}$, $Y^n = \{Y_m, m\leqslant n\}$.

如果（13.9）中的“ = ”换为“⩽”后仍成立，则X_n是**上鞅**；如果“ = ”换为“⩾”后仍成立，则X_n是**下鞅**. ◊

在很多情况下，我们并不指明随机变量Y_n，而只是说X_n是一个鞅、上鞅或下鞅. 如果X_n是一个鞅，那么

$$E(X_n) = E(X_0), \forall n \geqslant 0.$$

事实上，由条件期望的重期望公式（见定理 7.7）得到$E(X_n) = E(E[X_n|X_0, Y_0])$.

13.8.2 示例

下面来看一些鞅的例子.

1. 随机游走

设$\{Z_n, n\geqslant 0\}$为期望为零的独立随机变量. 那么，对于$n\geqslant 0$，$X_n := Z_0+\cdots+Z_n$是一个鞅，且

$$E[X_{n+1}|X^n] = E[Z_0+\cdots+Z_n+Z_{n+1}|Z_0, \cdots, Z_n] = Z_0+\cdots+Z_n = X_n.$$

注意到，若$E(Z_n)\leqslant 0$，那么X_n是一个上鞅；如果$E(Z_n)\geqslant 0$，那么X_n是一个下鞅.

2. 乘积

令$\{Z_n, n\geqslant 0\}$为均值为 1 的独立随机变量. 那么，对于$n\geqslant 0$，$X_n := Z_0\times\cdots\times Z_n$是一个鞅. 事实上，

$$E[X_{n+1}|X_n] = E[Z_0\times\cdots\times Z_n\times Z_{n+1}|Z_0, \cdots, Z_n] = Z_0\times\cdots\times Z_n = X_n.$$

如果对于所有n有$Z_n\geqslant 0$且$E(Z_n)\leqslant 1$，那么X_n是一个上鞅. 同理，如果对于所有n有$Z_n\geqslant 0$且$E(Z_n)\geqslant 1$，那么X_n是一个下鞅.

3. 分支过程

对于$m\geqslant 1$和$n\geqslant 0$，令X_m^n为与X独立同分布的随机变量. 其中X取值在$Z_+ := \{0, 1, 2\cdots\}$上且期望为μ. 则分支过程的定义为$Y_0 = 1$且

$$Y_{n+1} = \sum_{m=1}^{Y_n} X_m^n, n \geqslant 0.$$

该定义可以理解为在一个家族的第n代有Y_n个成员，其中第m个成员有X_m^n个子女.

根据定义，可以看到

$$Z_n = \mu^{-n} Y_n, n \geqslant 0$$

是一个鞅，因为

$$E[Y_{n+1} \mid Y_0, \cdots, Y_n] = Y_n \mu,$$

所以，

$$E[Z_{n+1} \mid Z_0, \cdots, Z_n] = E[\mu^{-(n+1)} Y_{n+1} \mid Y_0, \cdots, Y_n] = \mu^{-n} Y_n = Z_n.$$

设 $f(s) = E(e^s X)$且记 q 为 $q = f(q)$的最小非负解，则以下的过程为鞅：

$$W_n = q^{Z_n}, n \geqslant 1.$$

证明　留作练习. ♣

4. 杜布鞅

令$\{X_n, n = 1, \cdots, N\}$为随机变量且 $Y = f(X_1, \cdots, X_N)$，其中 f 是一个有界的可测实值函数. 那么，

$$Z_n := E[Y \mid X^n], n = 0, \cdots, N$$

是一个鞅（由条件期望的重期望性质，见定理 7.7），叫作*杜布鞅*. 下面是两个例子.

(1) 把 N 个球扔到 M 个桶中，令 Y 为某个抛掷过程的函数，如空桶的个数，最大负荷量，次大负荷量的桶数，等等. 令 X_n 为第 n 个球所在桶的编号，那么 $Z_n = E[Y|X_n]$ 是一个鞅.

(2) 假设一个桶中有 r 个红球，b 个蓝球. 我们从桶中取球（取出的球不放回）：拿出红球的数目是多少？令 X_n 为表示第 n 个球是否为红色的示性函数，令 $Y = X_1 + \cdots + X_n$ 为红球个数. 那么 Z_n 是一个鞅.

5. 你无法赢过庄家

为了研究收敛性，我们首先解释鞅的一个重要性质：在公平的投机游戏中没有必胜的法门.

定理 13.12　（你永远赢不了）

设 X_n 为对应于$\{X_n, Z_n, n \geqslant 0\}$的鞅，而 V_n 为关于(X_n, Z_n)的有界函数. 那么，

$$Y_n = \sum_{m=0}^{n} V_{m-1}\left(X_m - X_{m-1}\right), n \geqslant 1, \tag{13.10}$$

是一个鞅 $Y_0 := 0$.

证明

我们有

$$
\begin{aligned}
&E[Y_n - Y_{n-1} \mid X^{n-1}, Z^{n-1}] \\
&\quad = E[V_{n-1}(X_n - X_{n-1}) \mid X^{n-1}, Z^{n-1}] \\
&\quad = V_{n-1}E[X_n - X_{n-1} \mid X^{n-1}, Z^{n-1}] = 0.
\end{aligned}
$$

♣

式（13.10）可以这么理解：X_m表示你在第 m 轮时的财富．假设你在每轮下注时都参考过去的胜负信息(X^{m-1}, Z^{m-1})，并按照预期的盈利比 X_m-X_{m-1}下注．记 V_{m-1}为赌注大小．于是 Y_n代表了你在第 n 轮时的财富．上面的结果表明，即便实时调整赌注，你的最终获利也仍然是一个鞅．

6. 停时

玩投机游戏时，可以在输赢几次之后决定不再继续游戏．这种决定不以对下次输赢的预测为依据．这就意味着不能说："知道最后三局会输我才决定不玩的．"因此，随机停止时间 τ 使得事件$\{\tau\leqslant n\}$只取决于时刻 n 之前的信息．我们也称这样的随机时间为停时．

定义 13.13　停止时间

如果随机变量 τ 满足在$\{0, 1, 2, \cdots\}$中取值且存在函数 ϕ_n 使得

$$P[\tau \leqslant n \mid X_m, Y_m, m \geqslant 0] = \phi_n(X^n, Y^n), \forall n \geqslant 0$$

成立，则称 τ 是序列$\{X_n, Y_n, n\geqslant 0\}$的停止时间．　◊

举个例子，

$$\tau = \min\{n \geqslant 0 \mid (X_n, Y_n) \in \mathcal{A}\}$$

就是序列$\{X_n, Y_n, n\geqslant 0\}$的停止时间，其中 $\mathcal{A}$ 是一个在$\Re^2$上的集合．你想要停止的时候，要么已经一无所有，要么财富已经超过 1 千美元．

现在大家也许会想是不是有可能通过选择最佳的停止策略来让自己在游戏中的财富增加？事实证明，这是做不到的．如下面的定理所示．

定理 13.14　可选择的停止

设$\{X_n, n\geqslant 0\}$为鞅，τ 是对应于$\{X_n, Y_n, n\geqslant 0\}$的停时．那么，

$$E[X_{\tau\wedge n} \mid X_0, Y_0] = X_0.$$

在这一定理的陈述中，对于随机时间 σ，定义当 $\sigma = n$ 时，$X_\sigma := X_n$.

证明

注意到，如果玩家每次选择的赌注为$V_m = 1\{\tau \wedge n > m\}$，即在时间$\tau \wedge n$之前一直下赌注 1，由式（13.10）可以看到 $X_{\tau\wedge n}$ 是该玩家在时刻 n 的财富 Y_n．由于

$1\{\tau \wedge n > m\} = 1 - \{\tau \wedge n \leqslant m\} = \phi(X^m, Y^m)$，玩家最终得到的财富是一个鞅. ♣

读者也许已经注意到，限制上述定理中 $\tau \wedge n$ 的大小十分重要. 比如，设 X_n 对应于前面讨论过的随机游走，其中 $P(Z_n = 1) = P(Z_n = -1) = 0.5$. 如果定义 $\tau = \min\{n \geqslant 0 | X_n = 10\}$，则 τ 是有限的（详见定理 13.2 下的评述）. 因此，$X_\tau = 10$，

$$E[X_\tau \mid X_0 = 0] = 10 \neq X_0.$$

然而，如果我们界定停时，则该定理指出

$$E[X_{\tau \wedge n} \mid X_0 = 0] = 0. \tag{13.11}$$

这一结果值得大家进一步思考.

也许有人会对式（13.11）左边取极限：当 $n \to \infty$时，因为 τ 是有限的，所以

$$\lim_{n \to \infty} X_{\tau \wedge n} = X_\tau = 10,$$

大家可能会得出结论，式（13.11）的左边趋向 10，和式（13.11）矛盾了. 这里的问题在于取极限和取期望两个操作不能互换，因为随机变量 $X_{\tau \wedge n}$ 不是有界的. 但是，假如它们是有界的由控制收敛定理，可以得出 $E[X_\tau | X_0] = X_0$. 我们将这一观察结果总结为下面的定理.

定理 13.15　可选择的停止—2

设 $\{X_n, n \geqslant 0\}$ 为鞅，τ 是对应于 $\{X_n, Y_n, n \geqslant 0\}$ 的停时. 假设存在随机变量 V，满足 $E(V) < \infty$ 以及 $|X_n| \leqslant V$. 那么，

$$E[X_\tau \mid X_0, Y_0] = X_0.$$

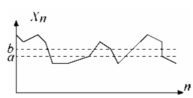


图 13-14　如果 X_n 不收敛，则存在一些有理数 $a < b$ 使得 X_n 穿过区间$[a, b]$无限多次

7. L^1 有界鞅

L^1 有界鞅不可能在区间$[a, b]$上穿越无限多次. 因为如果这样的话，那么玩家只要在向上穿过区间时下赌注 1，并在向下穿过区间时下赌注 0，就可以无穷地增加财富. 我们马上就能看到这不可能发生. 因此，这个鞅一定收敛.（注意，如果该鞅不是 L^1 有界的，那么该结果不成立，如随机游走例子所示.）

定理 13.16　L^1 有界鞅的收敛性

设 $\{X_n, n \geqslant 0\}$ 为鞅，且 $E(|X_n|) \leqslant K$ 对于所有 n 成立. 那么，X_n 几乎处处收敛于一个有限随机变量 X_∞.

证明

考虑区间[a, b]．我们现在说明 X_n 不能向上穿过该区间无限次（见图 13-14）．我们在 X_n 向上时下赌注 1；在其向下时下赌注 0．也就是说，首先等到 X_n 第一次小于 a，在后面每一步下赌注 1 直到 $X_n > b$；继而停止下注直到 X_n 再次小于 a．然后一直重复这一步骤．如果 X_m 在时间 n 之前穿过区间 U_n 次，玩家的财富至少有$(b–a)U_n–(X_n–a)$．这是因为，每次向上穿过区间时玩家的收益最少，为 $b—a$．而在最后几步，如果在玩家开始重新下注后，X_n 再也没有向上穿过 b，那么玩家最多输掉 $X_n—a$．但是，因为 Y_n 是一个鞅，所以

$$E(Y_n) = Y_0 \geqslant (b-a)E(U_n) - E(X_n - a) \geqslant (b-a)E(U_n) - K + a.$$

这说明 $E(U_n) \leqslant B = (K + Y_0 - a)/(b-a) < \infty$．令 $n\to\infty$，由于 $U_n\uparrow U$，其中 U 是向上穿过区间[a, b]的总数，由单调收敛定理得 $E(U)\leqslant B$．因此，U 是有限的，X_n 不能向上无限次地穿越任意给定区间[a, b]．所以，其向上无限次穿越有限区间的概率为 0（因为这样的区间个数是可数的）．这意味着 X_n 一定收敛，不是收敛于$+\infty$、$-\infty$就是收敛到一个有限值．但是因为 $E(|X_n|)\leqslant K$，所以 X_n 收敛于$+\infty$或$-\infty$的概率为 0． ♣

下面的定理是定理 13.16 的直接推论，十分有用．我们在定理 10.2 中就运用这一结果证明了随机梯度投影算法的收敛性．

定理 13.17 L^2 有界鞅的收敛性

设 X_n 为 L^2 有界鞅，即 $E(X_n^2) \leqslant K^2$，$\forall n \geqslant 0$．那么存在某有界的随机变量 X_∞，使得 X_n 几乎处处收敛到 X_∞.

证明

由詹森不等式，我们有

$$E(|X_n|)^2 \leqslant E(X_n^2) \leqslant K^2,$$

所以，对于所有 n 都有 $E(|X_n|)\leqslant K$．定理结论可以由定理 13.16 得到． ♣

我们还可以类似地证明 $E(|X_n–X_\infty|^2)\to 0$.

13.8.3 大数定律

鞅收敛结论还可以用于证明强大数定律，如杜布[Doob，1953]所证．

定理 13.18 强大数定律

令$\{X_n, n\geqslant 1\}$为独立同分布随机变量，其中 $E(|X_n|) = K<\infty$且 $E(X_n) = \mu$. 那么，当 $n \to \infty$ 时，$\dfrac{X_1 + \cdots + X_n}{n} \to \mu$ 几乎处处收敛.

证明

令

$$S_n = X_1 + \cdots + X_n, n \geqslant 1.$$

注意，由对称性有

$$E[X_1 \mid S_n, S_{n+1}, \cdots] = \frac{1}{n} S_n =: Y_{-n}, \tag{13.12}$$

所以，

$$\begin{aligned} E[Y_{-n} \mid S_{n+1}, \cdots] &= E[E[X_1 \mid S_n, S_{n+1}, \cdots] \mid S_{n+1}, \cdots] \\ &= E[X_1 \mid S_{n+1}, \cdots] = Y_{-n-1}. \end{aligned}$$

所以，$\{\cdots, Y_{-n-2}, Y_{-n-1}, Y_{-n}, \cdots\}$是一个鞅（杜布鞅）. 和前面一样，这说明该鞅向上穿过区间$[a, b]$的次数 U_n 满足 $E(U_n) \leqslant B < \infty$. 如前所述，我们得出结论 $U := \lim U_n < \infty$几乎处处收敛. 因此，Y_n 几乎必然收敛于随机变量 $Y_{-\infty}$.

现在，由于

$$Y_{-\infty} = \lim_{n \to \infty} \frac{X_1 + \cdots + X_n}{n},$$

我们知道对于任意有限 n，$Y_{-\infty}$独立于$(X_1, \cdots, X_n)$. 事实上，这个极限不依赖于前 n 个随机变量的值. 因为 $Y_{-\infty}$是关于$\{X_n, n \geqslant 1\}$的函数，所以它必然和自己独立，或者说它是一个常量.

因为 $E(Y_\infty) = E(Y_1) = \mu$，我们得到 $Y_\infty = \mu$. ♣

13.8.4 沃尔德等式

鞅的一个重要应用是下面将要介绍的沃尔德等式. 设$\{X_n, n \geqslant 1\}$为独立同分布随机变量. 令 τ 为随机变量，独立于从$\{1, 2, \cdots\}$中取值的 X_n，并有 $E(\tau) < \infty$. 那么，

$$E(X_1 + \cdots + X_\tau) = E(\tau) E(X_1). \tag{13.13}$$

这个表达式即为沃尔德等式.

理解这一结果的关键在于注意到 $Y_n = X_1 + \cdots + X_n Y_1 - nE(X_1)$是一个鞅. 而且 τ 是停时，所以只要将 τ 替换为 $\tau \wedge n$ 就得到：

$$E(Y_{\tau \wedge n}) = E(Y_1) = 0,$$

若 $E(\tau) < \infty$，可以令 n 趋向于无穷从而得到结论.（比如，将 X_i 替换为 X_i^+，然后用单调收敛定理对 X_i^- 进行类似的运算，然后相减.）

13.9 小结

- 一般推断问题：给定 Y 推断 X，贝叶斯准则运用与否
- 充分统计量：$h(Y)$是 X 的充分统计量
- 无限马尔可夫链：正递归、零递归、瞬态
- 李雅普诺夫–福斯特准则
- 泊松过程：独立稳态增量
- 连续时间马尔可夫链：速率矩阵
- 二元对称信道的香农容量：典型序列和随机编码
- 界限：切尔诺夫不等式和詹森不等式
- 鞅和收敛性
- 强大数定理

重要方程与公式

推断问题	给定Y猜测X:最大似然估计，最大后验概率，假设检验	13.1节
充分统计量	$f_{Y\|X}[y\|x]=f(h(y),x)g(y)$	定义13.1
无限马尔可夫链	不可约 $\Rightarrow$瞬态的，零递归或正递归	定理13.2
泊松过程	在接下来的 ε 秒内，跳跃概率为 λ_ε	定理13.4
连续时间马尔可夫链	从i跳到j的速率为$Q(i,j)$	定理13.8
香农容量	可以用任意速率$R<C$可靠地传输	13.6节
二元对称信道的香农容量	$C=1+p\log_2(p)+(1-p)\log_2(1-p)$	式（13.2）
切尔诺夫不等式	$P(X>a)\leqslant E(\exp\{\theta(X-a)\}),\forall\theta\geqslant 0$	式（13.5）
詹森不等式	h是凸函数 $\Rightarrow E(h(X))\geqslant h(E(X))$	式（13.6）
鞅	增量期望为0	定义13.11
鞅的收敛性	如果是L^1或L^2有界的，那么几乎处处收敛于有限的随机变量	定理13.16
沃尔德不等式	$E(X_1+\cdots+X_\tau)=E(\tau)E(X_1)$	式（13.13）

13.10 参考资料

有关于马尔可夫链的理论，可以参见 Chung 的 *Markov Chains with Stationary Transition Probabilities*. 教材 *Performance Modeling and Design of Computer Systems: Queueing Theory in Action* 讲解了基本的排队论和其在计算机系统和运筹学上的应用.

强烈推荐 Bremaud 的 *Markov Chains: Gibbs Fields*, *Monte Carlo Simulation*, *and*

Queues．该书的讲解清晰明了并且具有应用的广度．信息论在教材 *Elements of Information Theory* 中有详细阐述．我有关鞅的知识大多来自于 Neveu 的 *Discrete Parameter Martingales* 一书．

13.11 练习

1. 假设 $y_1, \cdots, y_n$ 是符合 $N(\mu, \sigma^2)$的独立同分布随机变量的样本点．给定 $\sigma = 1$，估计 μ 的充分统计量．给定 $\mu = 1$，估计 σ 的充分统计量?

2. 顾客到达商店服从泊松过程，速率为 4（每小时）.

(a) 在一小时内恰有 3 名顾客到达的概率是多少?

(b) 在第一名顾客到达之前，需要至少 40 分钟的概率是多少?

3. 考虑两个独立的泊松过程，它们的参数分别为 λ_1 和 λ_2．这两个过程计量到达商店 1 和商店 2 的顾客数.

(a) 有顾客到达商店 1 时，没有顾客到达商店 2 的概率是多少?

(b) 在第一个小时恰有 6 名顾客到达两个商店的概率是多少?（到达两个商店的人数之和为 6.）

(c) 在已知 6 名顾客到过两个商店的前提下,这 6 人全部去商店 1 的概率是多少?

4. 考虑图 13-15 所示的连续时间马尔可夫链.

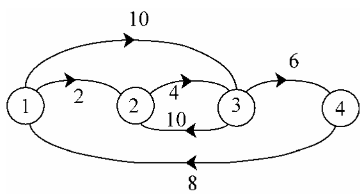


图 13-15 连续时间马尔可夫链

(a) 找出其稳态分布.

(b) 模拟该马尔可夫链，并证实其在状态 1 花费的时间比例收敛于 π(1).

5. 考虑一个离散时间排队系统．到达服从伯努利分布，参数为 λ．服务时间是独立同分布随机变量，且与到达时间独立．每次的服务时间 Z 从$\{1, 2, \cdots, K\}$中取值，并有 $E(Z) = 1/\mu$ 且 $\lambda < \mu$.

(a) 构建用于模拟这一排队问题的马尔可夫链. 其状态和转移概率各是什么?（提示：假设队伍最前面的任务仍然需要 z 个时间单位的服务．将 z 加入马尔可夫链的状态描述中.）

(b) 运用李雅普诺夫–福斯特准则证明该队列是稳定的，即马尔可夫链是正递归的.

6. 假设随机变量 X 从集合$\{1, 2, \cdots, K\}$中取值，使得 $\Pr(X_1 = k) = p_k > 0$ 且 $\sum_{k=1}^{K} p_k = 1$．假设 $X_1, X_2, \cdots, X_n$是 X 的 n 个独立同分布随机变量样本序列．

(a) 存在多少可能的序列？

(b) 当 n 很大时，存在多少典型序列？

(c) 找到一种情况，使得(a)和(b)的答案是相同的．

7. 令$\{N_t, t\geqslant 0\}$为参数为 λ 的泊松过程．令 S_n 表示第 n 个事件的时间．找出：

(a) S_n 的概率分布函数；

(b) $E[S_5]$；

(c) $E[S_4|N(1) = 2]$；

(d) $E[N(4)-N(2)|N(1) = 3]$．

8. 一队列到达服从泊松过程，参数为 λ．有两个并行工作的服务人员．当队列中至少有 2 名顾客时，有 2 名顾客同时接受服务．当只有 1 名顾客时，只有 1 位服务人员提供服务．服务时间是独立同分布随机变量，服从参数 $\mathrm{Exp}(\mu)$．

(a) 说明队列长度是马尔可夫链．

(b) 画出状态转移图．

(c) 找出使得队列是正递归的 μ 的最小值，并求解平衡方程．

9. 令$\{X_t, t\geqslant 0\}$为连续时间马尔可夫链，速率矩阵为 $Q = \{q(i, j)\}$．定义 $q(i) = \sum_{j\neq i} q(i, j)$．又令 $T_i = \inf\{t > 0 \mid X_t = i\}$ 且 $S_i = \inf\{t > 0 \mid X_t \neq i\}$．那么，以下哪些正确？

❑ $E[S_i \mid X_0 = i] = q(i)$．

❑ $P[T_i < T_j \mid X_0 = k] = q(k,i)/(q(k,i)+q(k,j))$，其中 i、j、k 各不相等．

❑ 若 $\alpha(k) = P[T_i < T_j \mid X_0 = k]$，那么对于 $k \in \{i, j\}$，$\alpha(k) = \sum_S \frac{q(k,s)}{q(k)}\alpha(S)$．

10. 一个连续时间队列有服从参数 λ 的泊松过程到达时间，且具备无限多的服务人员．对于众多顾客，服务人员可以并行工作．但他们之间互不合作，即 1 名顾客只能接受 1 名服务人员的服务．所以，当有 k 名顾客在队列中时，只有 k 名服务人员在工作．假设每名顾客的服务时间是独立同分布随机变量，服从参数为 μ 的指数分布．

(a) 说明队列长度是马尔可夫链，并画出该链的状态转移图．

(b) 证明对于有限的 λ 和 μ，马尔可夫链是正递归的，并找到其稳态分布．

11. 考虑一个参数为 $\lambda = 1$ 的泊松过程$\{N_t, t\geqslant 0\}$．令随机变量 S_i 表示第 i 个到达的时间．（提示：$f_{S_i}(x) = \dfrac{x^{i-1}\mathrm{e}^{-x}}{(i-1)!}1\{x \geqslant 0\}$．）

(a) 给定 $S_3 = s$，找出 S_1 和 S_2 的联合分布．

(b) 找出 $E[S_2|S_3 = s]$.

(c) 找出 $E[S_3|N_1 = 2]$.

12. 令 $S = \sum_{i=1}^{N} X_i$ 表示在 8 小时内从一台 ATM 取款的数额之和，其中：

(a) X_i是独立同分布随机变量，代表每个顾客的取款金额，且 $E[X_i] = 30$，$\mathrm{Var}[X_i] = 400$；

(b) N 是一个泊松随机变量，表示顾客总数，且 $E[N] = 80$；

求 $E[S]$和 $\mathrm{Var}[S]$.

13. 给定两个独立的泊松过程 M_t 和 N_t，其对应的速率分别为 λ 和 μ（$\lambda > \mu$）. 计算 $E(\tau)$，其中 $\tau = \max\{t \geqslant 0 \mid M_t \leqslant N_t + 5\}$（注意，这是最大值，不是最小值）.

14. 考虑一个队列,其到达服从泊松过程,参数为 λ . 服务时间为 1 个单位时间. 令 X_t 为时间 $t(t \geqslant 0)$时的队列长度.

(a) X_t 是一个马尔可夫链吗？证明或证伪之.

(b) 令 Y_n 是第 n 次（$n \geqslant 1$）离开队列后的队列长度. 证明 Y_n 是一个马尔可夫链，并画出其状态图.

(c) 证明当 $\lambda < 1$ 时，Y_n 是正递归的.

15. 假设一个队列系统的到达过程是参数为 λ 的泊松过程. 队列最多容纳 N 个顾客. 每个顾客所需的服务时间服从参数为 μ 的指数分布. 当 1 位顾客到达时，可以选择支付 c 元让其离开；如果 1 位顾客到达时队列已满，那么也需要支付 c 元. 你要决定什么时候接受新顾客，使得拒绝顾客的花费以及在队伍中的平均等待时间之和最小.

(a) 将这一问题建模成马尔可夫决策问题. 为了简化问题，考虑总的折扣代价：如果在时间 t 有 x_t 名顾客，那么在$[t, t+\varepsilon]$的等待代价为 $\mathrm{e}^{-\beta t} x_t \varepsilon$. 类似地，如果你在时刻 t 拒绝了一名顾客，那么代价为 $c\mathrm{e}^{-\beta t}$.

(b) 写出动态规划方程.

(c) 通过截断状态空间，运用 Matlab 求解该方程.

16. 计数过程 $\boldsymbol{N} := \{N_t, 0 \leqslant t \leqslant T\}$的定义如下：

给定 τ，$\{N_t, 0 \leqslant t \leqslant \tau\}$和$\{N_t - N_\tau, \tau \leqslant t \leqslant T\}$是独立的泊松过程，它们的参数分别为 λ_0 和 λ_1 . λ_0 和 λ_1 已知，且 $0 < \lambda_0 < \lambda_1$. 而且，τ 是指数分布的，其参数 μ（$\mu > 0$）已知.

(a) 给定 $\boldsymbol{N}$，求 τ 的 MLE.

(b) 给定 $\boldsymbol{N}$，求 τ 的 MAP.

17. 图 13-16 展现了一个系统. 在该系统中，输出源根据一个转移参数给定的连续时间马尔可夫链，在 ON 和 OFF2 之间相互转换. 当源处于 ON 时，它送出速率为

2 的水流进入队列．当源处于 OFF 时，则不送出水流．队列总以恒定速度 1 排水．令 X_t 为时间 $t \geqslant 0$ 时队列中的水量．

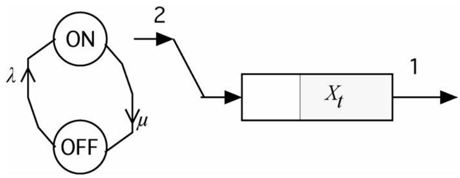


图 13-16　系统

(a) 画出这一随机过程$\{X_t, t \geqslant 0\}$的典型轨迹．

(b) 直观来看，λ 和 μ 需满足什么条件才能使得该队列稳定？

(c) 过程$\{X_t, t \geqslant 0\}$是马尔可夫过程吗？

18. $\{N_t, t \geqslant 0\}$为参数为 λ 的泊松过程，其中 λ 服从参数为 μ（$\mu > 0$）的指数分布．

(a) 求 $MLE[\lambda | N_s, 0 \leqslant s \leqslant t]$.

(b) 求 $MAP[\lambda | N_s, 0 \leqslant s \leqslant t]$.

(c) 给定$[N_s, 0 \leqslant s \leqslant t]$，$\lambda$ 的充分统计量是什么？

(d) 假设 λ 在[5, 10]中取值，不服从指数分布．求时间 t 的估计值，使得 λ 的估计误差不大于 5%的概率为 95%.

19. 考虑两个平行的离散时间队列，其到达过程分别为参数为 λ_1 和 λ_2 的伯努利过程．顾客的服务时间分别服从参数为 μ_1 和 μ_2 的几何分布．由于只有 1 名服务员，同一时间只能为队伍 1 或队伍 2 提供服务．考虑以下的调度策略：如果 $\mu_1 Q_1(n) > \mu_2 Q_2(n)$，那么服务员在时间 n 服务队伍 1，否则服务队伍 2，其中 $Q_1(n)$和 $Q_2(n)$是时刻 n 的队列长度．运用李雅普诺夫函数 $V(Q_1(n), Q_2(n)) = Q_1^2(n) + Q_2^2(n)$ 证明：如果 $\lambda_1 / \mu_1 + \lambda_2 / \mu_2 < 1$，则队列是稳定的．这一调度策略叫作最大权重或回压策略．

附录A

概率论基础知识

> **主题：** 总体框架，条件概率，贝叶斯定理，独立事件，期望，概率密度函数，累计分布函数，随机变量的函数，相关，方差，联合概率密度函数的转换

A.1 总体框架

概率论的基本模型看起来也许有些抽象、难以理解. 但事实上，概率论把所有的关键概念统一起来，形成了非常清晰的理论体系. 在研究书中的实际例子时，读者应当把整体的知识框架牢记在心.

A.1.1 概率空间

在描述随机性实验时，首先要用集合Ω来明确地表述出所有可能的结果. 这个集合叫作样本空间. 比如说，抛一枚硬币时，样本空间为$\Omega=\{H,T\}$（这里的H表示正面向上，而T表示反面向上）；掷一颗骰子时，样本空间为$\Omega=\{1,2,3,4,5,6\}$；测量电压时，样本空间可能是$\Omega=\Re=(-\infty,+\infty)$；等等.

确定样本空间以后，我们希望知道随机性实验结果落在Ω的某一个子集中的概率. 也就是说，对于子集$A\subset\Omega$，我们希望计算出概率$P(A)\in[0,1]$，这个数值能够代表随机性实验结果落在子集A中的可能性. 举例来说，在掷骰子的时候，用$|A|$来表示子集A中含有的元素个数，那么实验结果落在集合A中的概率为$P(A)=|A|/6$. 在测量电压时，测量值取得某一个特定数值的概率一般为0，但测量值的绝对值小于15的概率则可能是95%. 这就是定义概率时针对于集合，而不是针对于特定某个结果的原因.

当然，在定义Ω子集的概率时需要遵循一些规则. 比如说，如果$A\subset B$，那么一定有$P(A)\leqslant P(B)$. 同时，我们有$P(\Omega)=1$. 另外，假如集合 A、B 没有交集，那么

$P(A\bigcup B)=P(A)+P(B)$．最后，为了能够用简单集合来近似复杂集合，我们还要求：如 $A_1\subset A_2\subset A_3\subset\cdots$ 且 $A=\bigcup_n A_n$，则 $P(A_n)\to P(A)$；同样地，如果 $A_1\supset A_2\supset A_3\supset\cdots$ 且 $A=\bigcap_n A_n$，则 $P(A_n)\to P(A)$．

从上述性质可以推出如下结论．

定理 A.1　博雷尔–坎特利引理

令事件 A_n 满足

$$\sum_{n=1}^{\infty}P(A_n)<\infty,$$

则有

$$P(A_n,\text{i.o.})=0.$$

这里，$\{A_n,\text{i.o.}\}$ 代表有无穷多个事件 A_n 发生．□

证明　首先，我们注意到：

$$\{A_n,\text{i.o.}\}=\bigcap_n B_n,$$

这里的 $B_n=\bigcup_{m\geqslant n}A_m$ 是一串递减的集合．事实上，当且仅当对于所有的 n 都有某个事件 A_m 在 $m\geqslant n$ 时发生，A_n 会出现无穷多次．

因此，

$$P(A_n,\text{i.o.})=\lim_{n\to\infty}P(B_n)$$

并且 $P(B_n)\leqslant\sum_{m=n}^{\infty}P(A_m)$，因此[①]

$$P(B_n)\to 0\text{，}\quad n\to\infty. \qquad ♣$$

在这里需要说明，一般情况下无法严格定义 Ω 所有子集的概率，不过这并不影响概率的应用．在专业术语中，我们把 Ω 中能够在其上严格定义概率的子集称为事件．

A.1.2　独立事件

对于两个事件 A 和 B，如果 $P(A\bigcap B)=P(A)P(B)$，那么称 A 和 B 是独立的．比如同时投掷两枚骰子，得到的结果是 $(a,b)\in\{1,2,\cdots,6\}^2$．$a$ 表示第一枚骰子的点数，b 表示第二枚骰子的点数．现在考虑两个事件：事件“第一枚骰子的点数在集合 $\{2,4,5\}$ 中”对应结果的集合是 $A=\{2,4,5\}\times\{1,\cdots,6\}$；事件“第二枚骰子的点数在集合 $\{2,4\}$ 中”

① 如果非负的数列 a_n 满足 $\sum_{n=0}^{\infty}a_n<\infty$，则 $n\to\infty$ 时一定有 $\sum_{m=n}^{\infty}a_m\to 0$

对应结果的集合是 $B=\{1,\cdots,6\}\times\{2,4\}$．根据定义可以判断 A 和 B 独立，因为 $P(A)=18/36$，$P(B)=12/36$，而且 $P(A\cap B)=6/36$．

一个更复杂一些的概念是相互独立．如果 $\{A_j,\ j\in J\}$ 满足

$$P(\bigcap_{j\in K}A_j)=\prod_{j\in K}P(A_j),\ \forall \text{有限集}\ K\subset J,$$

则称事件 $\{A_j,\ j\in J\}$ 相互独立．我们可以很容易地构造出两两独立但不相互独立的集合．比如，假设 $\Omega=\{1,2,3,4\}$ 且 4 个结果出现的可能性相等．令 $A=\{1,2\}$，$B=\{1,3\}$ 以及 $C=\{1,4\}$，则很容易验证这 3 个子集两两独立但不相互独立，因为 $P(A\cap B\cap C)=1/4\neq P(A)P(B)P(C)$．

A.1.3 随机变量

随机变量 X 是一个从样本空间到实数集的函数 $X:\Omega\to\Re$．通过这个概念，可以将随机试验中每一个可能的结果与一个实数 $X(\omega)$ 联系起来．

例如在抛硬币的试验中，样本空间为 $\Omega=\{H,T\}$．可以将随机变量 X 定义为 $X(H)=1$，$X(T)=0$．

接下来，对于实数集 $B\subset\Re$，采用记号 $P(X\in B)=P(X^{-1}(B))$，其中

$$X^{-1}(B):=\{\omega\in\Omega\,|\,\mathrm{X}(\omega)\in B\}.$$

这个记号表达的意思很直观：随机变量 $X\in B$ 的概率等于所有使得 $X(\omega)\in B$ 的实验结果 ω 的概率．

另外，特别定义累积分布函数为 $F_X(x)=P(X\in(-\infty,x])=:P(X\leqslant x)$．

图 A-1 是一个随机变量的示意图．

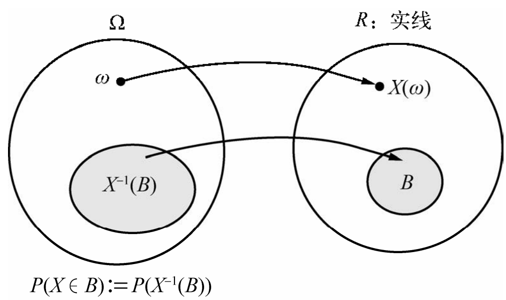


图 A-1 随机试验可以用结果集 Ω 即样本空间进行描述．每个 Ω 的子集都有一个相应的概率，而每个随机变量则是一个随机试验结果 ω 的实值函数

A.1.4 随机向量

随机向量 $\boldsymbol{X}=(X_1,\cdots,X_n)'$ 的每一个分量都是定义在同一个概率空间中的随机变

量．也就是说，它是一个从采样空间到 $\Re^n$ 的函数 $\boldsymbol{X}:\Omega\to\Re^n$．按如下方式定义联合累积分布函数：

$$F_{\boldsymbol{X}}(\boldsymbol{x})=P(X_1\leqslant x_1,\cdots,X_n\leqslant x_n),\ \boldsymbol{x}\in\Re^n.$$

A.2　离散型随机变量

A.2.1　定义

我们将离散型随机变量 X 定义为一系列离散可能的取值及它们对应的概率：

$$X\equiv\{(x_n,p_n),n=1,2,\cdots,N\}.\qquad\text{(A.1)}$$

C.2

这里的 x_n 是实数，p_n 是正实数且和为 1. 根据定义，p_n 是随机变量 X 取值为 x_n 的概率，记为

$$p_n=P(X=x_n),n=1,\cdots,N.$$

N 的取值可以是无穷大．这一系列的 p_n 被称为随机变量 X 对应的概率质量函数．

举个例子，

$$V\equiv\{(1,0.1),(2,0.3),(3,0.6)\}$$

是一个随机变量，它有 3 个可能的取值(1, 2, 3)，相应的概率为(0.1, 0.3, 0.6).

离散型随机变量与概率论总体框架的联系如下：有一个概率空间和一个函数 $X:\Omega\to\Re$，它仅在集合 $\{x_1,\cdots,x_N\}$ 中取值，并且满足

$$P(X=x_n)=P(\{\omega\in\Omega\,|\,X(\omega)=x_n\})=p_n.$$

对某一个随机变量 X 感兴趣时，应当关注它可能的取值及其对应的概率，即随机变量的概率质量函数，而不必关注产生这个随机变量所需要的随机试验.

A.2.2　期望

随机变量 X 的期望值（或平均值）记为 $E(X)$，定义为

$$E(X)=\sum_{n=1}^{N}x_np_n.$$



图 A-2　随机变量的期望值

在上文的例子中，

$$E(V)=1\times0.1+2\times0.3+3\times0.6=2.5.$$

下面是另一个经常用到的例子：如果 A 是样本空间 Ω 中的一个事件，令 $X(\omega)=1\{\omega\in A\}$，则称随机变量 X 是事件 A 的指示函数. 在这种情况下，事件 A 以 $P(A)$ 的概率发生，此时 $X=1$；其余情况下 $X=0$. 所以 $E(X)=P(A)$.

当 N 为无穷大时，上述定义依然成立，除非在 $E(X)$求和时所有正项之和与所有负项之和都是无穷大. 这种情况下，X 的期望值不存在.

A.2.3 随机变量的函数

考虑一个离散型随机变量和函数 $h:\Re\to\Re$. 则 $h(X)$定义了一个新的随机变量，其取值和概率值为

$$\{(h(x_n),p_n),n=1,\cdots,N\}.$$

注意 $h(x_n)$ 对不同的 x_n 取值可能是相同的，为了与概率质量函数的定义相符，我们需要将取值相同的项合并，并将其概率相加.

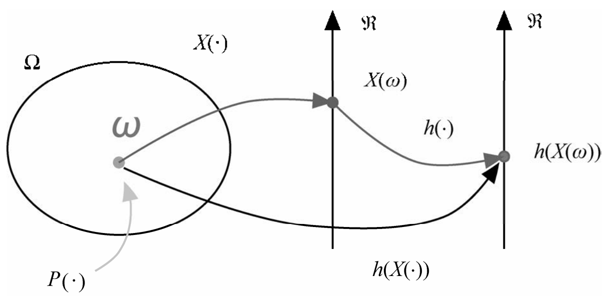


图 A-3 随机变量的函数

例如，如果 $h(1)=h(2)=10$，$h(3)=15$，那么

$$h(V)\equiv\{(10,0.4),(15,0.6)\},$$

这里将 $h(1)$ 和 $h(2)$ 两项合并了，因为它们的值都是 10.

所以，

$$E(h(V))=10\times0.4+15\times0.6=13.$$

注意到

$$E(h(V))=\sum_{n=1}^{N}h(v_n)p_n.$$

这是因为

$$\sum_{n=1}^{3} h(v_n)p_n = h(1)0.1 + h(2)0.3 + h(3)0.6$$
$$= 10\times 0.1 + 10\times 0.3 + 15\times 0.6$$
$$= 10\times(0.1+0.3) + 15\times 0.6,$$

与之前的表达式相符.

下面将上述观察总结成定理.

定理 A.2 随机变量函数的期望

假设随机变量 X 的概率质量函数为 $\{(x_n, p_n), n=1,\cdots,N\}$ 且 $h:\Re\to\Re$ 是一个函数. 则

$$E(h(X)) = \sum_{n=1}^{N} h(\mathrm{x}_n)p_n. \qquad \square$$

A.2.4 非负的随机变量

如果随机变量 X 的所有可能取值 x_n 都是非负的，则称 X 是非负的，写作 $X\geqslant 0$. 注意到

$$若 X\geqslant 0 且 E(X)=0，则 P(X=0)=1 .$$

同样,

$$若 X\geqslant 0 且 E(X)<\infty，则 P(X<\infty)=1 .$$

A.2.5 期望的线性性质

考虑函数 $h_1:\Re\to\Re$ 和 $h_2:\Re\to\Re$，定义 $h_1(X)+h_2(X)$ 如下：

$$h_1(X)+h_2(X) \equiv \{(h_1(x_n)+h_2(x_n), p_n), n=1,\ldots,N\}.$$

则有

$$E(h_1(X)+h_2(X)) = \sum_{n=1}^{N}(h_1(x_n)+h_2(x_n))p_n.$$

将右侧各项重组，有

$$E(h_1(X)+h_2(X)) = E(h_1(X)) + E(h_2(X)),$$

即期望具有线性的性质.

A.2.6 方差与变异系数

定义随机变量 X 的方差 $\mathrm{var}(X)$ 为

$$\mathrm{var}(X) = E((X-E(X))^2),$$

根据期望的线性性质可以得到

$$\begin{aligned}\operatorname{var}(X) &= E(X^2 - 2XE(X) + E(X)^2) \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 = E(X^2) - E(X)^2.\end{aligned}$$

对于式（A.1）中的例子，有

$$\operatorname{var}(V) = E(V^2) - E(V)^2 = 1^2 \times 0.1 + 2^2 \times 0.3 + 3^2 \times 0.6 - (2.5)^2 = 0.45.$$

定义随机变量 X 的*标准差* σ_X 为其方差的平方根，即

$$\sigma_X := \sqrt{\operatorname{var}(X)}.$$

注意到，如果随机变量 W 以相同的概率取值 $E(X)-\sigma_X$ 或 $E(X)+\sigma_X$，那么 $E(W)=E(X)$ 并且 $\operatorname{var}(W)=\operatorname{var}(X)$．从这个角度来看，$\sigma_X$ 是对期望值偏差的度量．

读者可以轻松验证，使得 $E((X-a)^2)$ 取最小值的 a 满足 $a=E(X)$．因此，期望也叫作随机变量 X 在最小二乘意义下的估计值．



图 A-4　方差使随机性变得有趣

A.2.7　重要的离散型随机变量

这里介绍几个重要的离散型随机变量．

1. 伯努利随机变量

如果

$$X = \{(0, 1-p), (1, p)\},$$

即如果

$$P(X=0) = 1-p \text{ 且 } P(X=1) = p,$$

那么 X 是关于参数 $p \in [0,1]$ 的伯努利随机变量，记为 $X =_D B(p)$．

请读者验证 $E(X)=p$ 以及 $\operatorname{var}(X)=p(1-p)$．这个随机变量能够用于为单次抛硬币问题建模，其中 1 代表正面向上，0 代表反面向上．



图 A-5 伯努利随机变量用于为单次抛硬币问题建模

2. 几何随机变量

如果

$$P(X=n)=(1-p)^{n-1}p, n\geqslant 1,$$

则称随机变量 X 关于参数 $p\in[0,1]$ 服从几何分布，记为 $X=_D G(p)$.

请读者验证 $E(X)=1/p$ 以及 $\operatorname{var}(X)=p(1-p)^{-2}$. 这个随机变量也可以用于为抛硬币问题建模，描述了第一次出现反面时抛掷的次数.（有时，$X-1$ 也被称为在 $\{0,1,\cdots\}$ 上的几何随机变量. 为了避免混淆，本书将 X 定义在集合 $\{1,2,\cdots\}$ 上.）



图 A-6 几何随机变量用于为第一次出现反面时的硬币抛掷次数建模

3. 二项随机变量

如果

$$P(X=n)=\binom{N}{n}p^n(1-p)^{N-n}, n=0,\cdots,N,$$

则称随机变量 X 为关于参数 N 和 p 的二项随机变量，记为 $X=_D B(N,p)$. 这里

$$\binom{N}{n}=\frac{N!}{(N-n)!n!}.$$

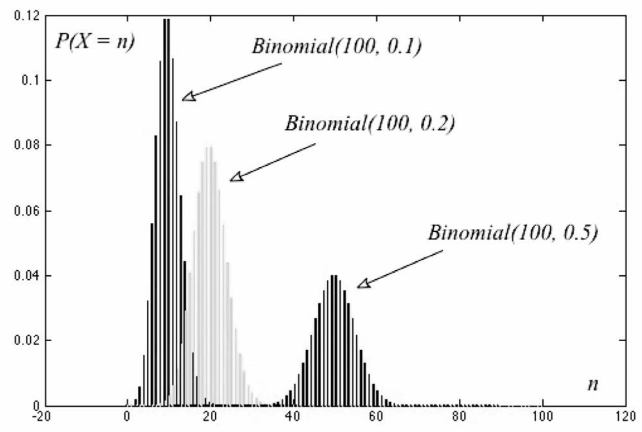


图 A-7 二项分布 B（100, P）对应的概率质量函数，其中 P 的取值分别为 0.1、0.2 和 0.5

请读者验证 $E(X)=Np$ 以及 $\mathrm{var}(X)=Np(1-p)$. 这个随机变量描述了抛掷 N 次硬币之后正面朝上的次数，它是 N 个相互独立的伯努利随机变量之和.

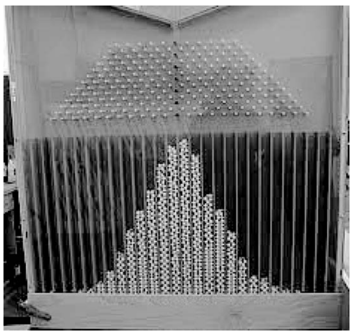


图 A-8 二项分布是伯努利随机变量之和

4. 泊松随机变量

如果

$$P(X=n)=\frac{\lambda^n}{n!}\mathrm{e}^{-\lambda}, n \geqslant 0, \tag{A.3}$$

则称随机变量 X 是关于参数 λ 的泊松随机变量，记为 $X =_D P(\lambda)$.

请读者验证 $E(X)=\lambda$ 以及 $\mathrm{var}(X)=\lambda$. 这个随机变量可以用于描述一个人一天收到的短信数量.

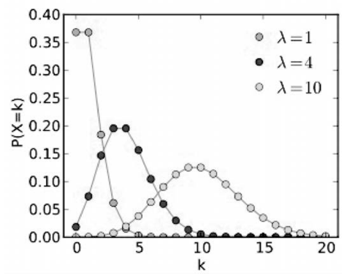


图 A-9　泊松概率质量函数，来源：维基百科（另见彩插）

A.3　一般的随机变量

并不是所有随机变量的取值都是离散的．对于电话线两端的电压值、风速、温度以及下一位顾客来到收银台需要的时间等变量，其取值可能是连续集合．

实际生活中，人们经常用有限位数的数字来近似表示一个数值．例如，以摄氏度为单位测量温度时，通常舍去小数部分，同时给出误差的上下限．因此，具有任意精确度要求的系统都可以通过离散随机变量来描述．然而，这种离散化带有很强的人工痕迹，有时甚至会使问题复杂化．例如，在牛顿方程 $F=ma$ 中，用离散变量来描述 $a=\mathrm{d}v(t)/\mathrm{d}t$ 是非常奇怪的，因为离散的速度值并不存在导数．因此，虽然计算机基于离散变量进行计算，但是在分析与推导时采用一般的变量更为合适．不能否认的是，离散值近似的方法依然发挥着重要的作用．

首先从一般随机变量的定义说起．

定义 A.3　累积分布函数与概率密度函数

设 X 是一个随机变量．

(a) X 的**累积分布函数** $F_X(x)$ 定义为：

$$F_X(x)=P(X\leqslant x), x\in\Re.$$

(b) 当 $F(x)$的导数存在时，X 的**概率密度函数**定义为

$$f_X(x)=\frac{\mathrm{d}}{\mathrm{d}x}F_X(x).$$ ◊

这里我们注意到，若 $a<b$，则

$$P(a<X\leqslant b)=P(X\leqslant b)-P(X\leqslant a)=F_X(b)-F_X(a)=\int_a^b f_X(x)\,\mathrm{d}x,$$

最后一个表达式仅在导数存在时有意义．

如果概率密度函数存在，则

$$f_X(x)\mathrm{d}x = F_X(x+\mathrm{d}x) - F_X(x) = P(X \in (x, x+\mathrm{d}x)).$$

这条性质解释了“概率密度”的含义.

示例

例 1 $U[a,b]$

如果

$$f_X(x) = \frac{1}{b-a} 1\{a \leqslant x \leqslant b\},$$

则称 X 是在区间 $[a,b]$ 上均匀分布的随机变量（ $a<b$ ），记为 $X =_D U[a,b]$. 这时，我们有

$$F_X(x) = \max\{0, \min\{1, \frac{x-a}{b-a}\}\}.$$

图 A-10 表明了均匀分布随机变量 $U[a,b]$ 的概率密度函数与累积分布函数.

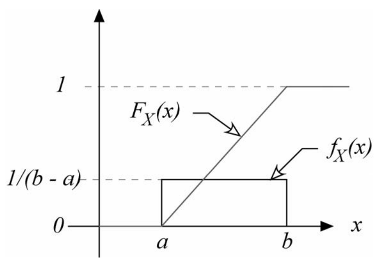


图 A-10 随机变量 $U[a, b]$ 的概率密度函数和累积分布函数

例 2 $Exp(\lambda)$

如果

$$f_X(x) = \lambda \mathrm{e}^{-\lambda x} 1\{x \geqslant 0\},$$

则称 X 是参数为 λ 的指数分布随机变量($\lambda > 0$)，记为 $X =_D \mathrm{Exp}(\lambda)$. 读者可以验证

$$F_X(x) = 1 - \exp\{-\lambda x\}, x \geqslant 0.$$

因此

$$P(X \geqslant x) = \exp\{-\lambda x\}, \forall x \geqslant 0.$$

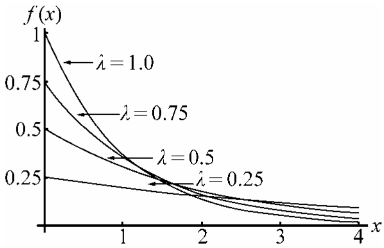


图 A-11 指数分布的密度（另见彩插）

如果希望更直观地理解以上内容，可以通过用一个定义在 $\{\cdots,-2\varepsilon,-\varepsilon,0,\varepsilon,2\varepsilon,\cdots\}$ 上的离散型随机变量 Y 来近似累积分布函数为 $F_X(\cdot)$ 的随机变量 X. 这时候，Y 的概率分布为：

$$P(Y=n\varepsilon)=F_X((n+1)\varepsilon)-F_X(n\varepsilon)=P(X\in(n\varepsilon,(n+1)\varepsilon]).$$

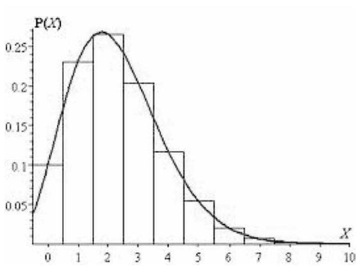


图 A-12 连续型随机变量的离散逼近

对于 Y 的函数 $h:\Re\to\Re$，我们有

$$h(Y)=\sum_n h(n\varepsilon)[F_X((n+1)\varepsilon)-F_X(n\varepsilon)]\approx\int_{-\infty}^{\infty}h(x)\mathrm{d}F_X(x).$$

其中最后一个式子为当 $\varepsilon\to 0$ 时求和的极限. 如果概率密度函数存在，上述公式则变成

$$h(Y)\approx\int_{-\infty}^{\infty}h(x)f_X(x)\mathrm{d}x.$$

如果 ε 取值足够小，那么 Y 作为 X 近似的误差就足够小. 因此，$E(h(Y))$ 会无限趋近于 $E(h(X))$. 将上述观察总结为如下定理.

定理 A.4 随机变量函数的期望

设 X 是一个累积分布函数为 $F_X(\cdot)$ 的随机变量，$h:\Re\to\Re$ 是一个函数，则

$$E(h(X))=\int_{-\infty}^{\infty}h(x)\mathrm{d}F_X(\mathrm{x}).$$

若 X 的概率密度函数 $f_X(\cdot)$ 存在，则

$$E(h(X)) = \int_{-\infty}^{\infty} h(x) f_X(x) \, \mathrm{d}\, x.$$

例如，若 $X =_D U[0,1]$，则 □

$$E(X^k) = \int_0^1 x^k \mathrm{d}x = \frac{1}{k+1}.$$

同时，X 的方差为：

$$\mathrm{var}(X) = E(X^2) - E(\mathrm{X})^2 = \frac{1}{3} - (\frac{1}{2})^2 = \frac{1}{12}.$$

再举一个例子．设 $X =_D \mathrm{Exp}(\lambda)$，则

$$E(X) = \int_0^{\infty} x\lambda \mathrm{e}^{-\lambda x} \mathrm{d}x = -\int_0^{\infty} x \mathrm{d}\mathrm{e}^{-\lambda x} = -[x\mathrm{e}^{-\lambda x}]_0^{\infty} + \int_0^{\infty} \mathrm{e}^{-\lambda x} \mathrm{d}x$$
$$= -\lambda^{-1}[\mathrm{e}^{-\lambda x}]_0^{\infty} = \lambda^{-1}.$$

同时，我们也有

$$E(X^2) = \int_0^{\infty} x^2 \lambda \mathrm{e}^{-\lambda x} \mathrm{d}x = -\int_0^{\infty} x^2 \mathrm{d}\mathrm{e}^{-\lambda x}$$
$$= -[x^2 \mathrm{e}^{-\lambda x}]_0^{\infty} + \int_0^{\infty} \mathrm{e}^{-\lambda x} \mathrm{d}x^2$$
$$= 2\int_0^{\infty} x \mathrm{e}^{-\lambda x} \mathrm{d}x = 2\lambda^{-2}.$$

因此，X 的方差为：

$$\mathrm{var}(X) = E(X^2) - E(\mathrm{X})^2 = 2\lambda^{-2} - (\lambda^{-1})^2 = \lambda^{-2}.$$

下面介绍一个容易混淆的例子．考虑这样的一个随机变量，它取值 0.3 的概率是 0.4，均匀分布在[0, 1]上的概率是 0.6．这相当于抛掷一枚得到正面概率为 0.4 的硬币．如果得到正面，则随机变量取值为 $X = 0.3$；如果得到反面，则 X 从[0, 1]中均匀取值．那么，

$$F_X(x) = P(X \leqslant x) = 0.4 \times 1\{x \leqslant 0.3\} + 0.6x, \; x \in [0,1].$$

对应的累积分布函数曲线如图 A-13 所示．现在可以通过引入狄拉克的冲激函数作为阶跃函数的导数来获得 $F_X(x)$的导数．

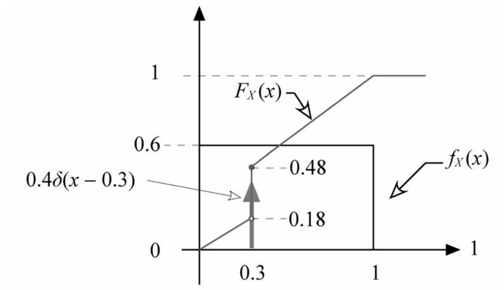


图 A-13　混合随机变量 X 的概率密度函数和累积分布函数（另见彩插）

对于这个随机变量，可以发现①

$$E(X^k)=\int_{-\infty}^{\infty}x^k f_X(\mathrm{d}x)$$
$$=\int_{-\infty}^{\infty}x^k 0.4\delta(x-0.3)\mathrm{d}x+\int_0^1 x^k 0.6\mathrm{d}x$$
$$=0.4(0.3)^k+0.6\frac{1}{k+1}.$$

于是我们得到

$$\mathrm{var}(X)=E(X^2)-E(X)^2$$
$$=0.4(0.3)^2+0.6\frac{1}{3}-[0.4(0.3)+0.6\frac{1}{2}]^2=0.0596.$$

还可以推出下面的重要定理.

定理 A.5　随机变量的期望为互补累积分布函数的积分

设 $X\geqslant 0$ 为非负的随机变量，则

$$E(X)=\int_0^{\infty}P(X>x)\mathrm{d}x.$$

证明 □

由

$$\frac{\mathrm{d}}{\mathrm{d}x}[u(x)v(x)]=u'(x)v(x)+u(x)v'(x)$$

推导出分部积分的公式：

$$\int_a^b u(x)\mathrm{d}v(x)=[u(x)v(x)]_a^b-\int_a^b v(x)\mathrm{d}u(x)\ .$$

① 狄拉克冲激函数的定义为 $\int_{-\infty}^{\infty}g(x)\delta(x-a)\mathrm{d}x=g(a)$，其中 $g(\cdot)$是在 a 点连续的任意函数.

根据分部积分公式得到

$$E(X) = \int_0^\infty x\mathrm{d}F_X(x) = -\int_0^\infty x\mathrm{d}(1 - F_X(x))$$
$$= -[xF_X(x)]_0^\infty + \int_0^\infty (1 - F_X(x))\mathrm{d}x = \int_0^\infty P(X > x)\mathrm{d}x.$$

函数 $P(X > x) = 1 - F_X(x)$ 被称为互补累积分布函数. ♣

例如，若 $X =_D Exp(\lambda)$，则

$$E(X) = \int_0^\infty P(X > x)\mathrm{d}x = \int_0^\infty \exp\{-\lambda x\}\,\mathrm{d}\,x = \frac{1}{\lambda}.$$

A.4 多个随机变量

人们常常会关注彼此关联的多个随机变量. 例如，一个人的身高与体重，发送机发出的电压值与接收机收到的电压值，排队的人数与需要等待的时间. 这些都是相互关联的随机变量.

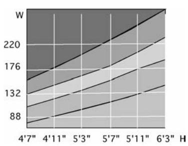


图 A-14 身高与体重是有关联的

要研究相互关联的随机变量，不能单独研究每个随机变量，而是需要全面的描述. 考虑下面的例子. 掷一枚骰子，设 $X = 1$ 表示朝上的点数是奇数，$X = 0$ 表示偶数. 同时，设 $Y = 1$ 表示结果在集合 $\{2, 3, 4\}$ 中，$Y = 0$ 表示结果在集合 $\{1, 5, 6\}$ 中. 注意 $P(X = 1) = P(X = 0) = 0.5$，$P(Y = 1) = P(Y = 0) = 0.5$. 因此，$X$、$Y$ 可以分别描述独立抛掷均匀硬币的结果. 然而，当我们把 X 和 Y 放在一起考虑时，(X, Y)就无法表示 2 次独立抛硬币的结果了. 比如，只有当点数为 3 时 $X = 1$ 且 $Y = 1$，这样的概率为 1/6. 如果 X、Y 表示 2 次独立抛均匀硬币的结果，则 $X = 1$ 且 $Y = 1$ 的概率应当是 1/4.

在离散情况下，我们通过列出所有可能的取值和相应的概率值来描述一对随机变量(X, Y)（见图 A-15）：

$$p_{i,j} = P(X = x_i, Y = y_j), \forall (i, j) \in \{1, \cdots, m\} \times \{1, \cdots, n\}$$

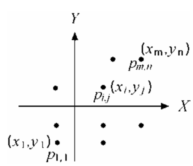


图 A-15 一对随机变量

这里的 $p_{i,j}$ 非负且和为 1，而 m、n 可以为无穷大．这样的描述确定了随机变量(X, Y)的联合概率质量函数．我们可以从联合概率质量函数中单独分离出 X 和 Y 的概率．比如，

$$P(X=x_i)=\sum_{j=1}^{n}P(X=x_i,Y=y_j)=\sum_{j=1}^{n}p_{i,j}.$$

对于两个随机变量 X 和 Y，如果

$$P(X=x,Y=y)=P(X=x)P(Y=y), \forall x,y,$$

则称 X、Y 是独立的．



图 A-16　什么是独立

在掷骰子的例子中，注意到

$$P(X=1,Y=1)=\frac{1}{6}\neq P(X=1)P(Y=1)=\frac{1}{4},$$

因此，X 和 Y 不是相互独立的．

对于函数 $h:\Re^2\to\Re$，定义

$$E(h(X,Y))=\sum_{k=1}^{N}h(x_k,y_k)p_k,$$

上式在求和不出现 $\infty-\infty$ 时成立．注意到，如果 $h(x,y)=h_1(x,y)+h_2(x,y)$，则

$$\begin{aligned}E(h(X,Y))&=\sum_{k=1}^{N}h(x_k,y_k)p_k\\&=\sum_{k=1}^{N}[h_1(x_k,y_k)+h_2(x_k,y_k)]p_k\\&=\sum_{k=1}^{N}h_1(x_k,y_k)p_k+\sum_{k=1}^{N}h_2(x_k,y_k)p_k\\&=E(h_1(X,Y))+E(h_2(X,Y)),\end{aligned}$$

因此期望具有线性性质．

我们定义 X 和 Y 的协方差为

$$\operatorname{cov}(X,Y)=E((X-E(X))(Y-E(Y))).$$

根据期望的线性性质，我们得到

$$\operatorname{cov}(X,Y)=E(XY-E(X)Y-XE(Y)+E(X)E(Y))=E(XY)-E(X)E(Y).$$

当$\operatorname{cov}(X,Y)=0$时，$X$和$Y$是不相关的；当$\operatorname{cov}(X,Y)>0$时，$X$和$Y$是正相关的；当$\operatorname{cov}(X,Y)<0$时，$X$和$Y$是负相关的.

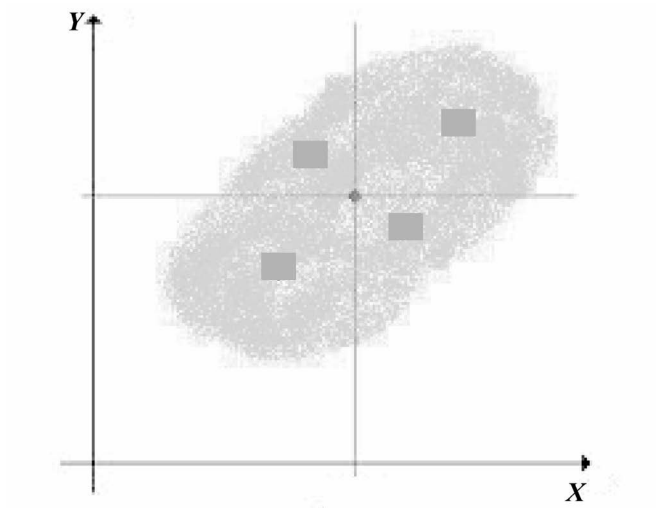


图 A-17　图中的随机变量为正相关：如果一个的值很大，另一个的值往往也很大

在前面掷骰子的例子中，

$$\operatorname{cov}(X,Y)=E(XY)-E(X)E(Y)=\frac{1}{6}-\frac{1}{4}<0,$$

因此X和Y是负相关的. 负相关性表明，当X大于平均值时，Y往往会小于平均值. 在我们的例子中，当$X=1$时出现的结果是奇数，而且$Y=0$的可能性要大于$Y=1$.

关于以上的内容有如下重要结论.

定理 A.6　相互独立的随机变量不相关

(a) 相互独立的随机变量不相关.

(b) 不相关的随机变量不一定相互独立.

(c) 不相关随机变量之和的方差等于这些随机变量的方差之和.

□

证明

(a) 设X和Y相互独立，则

$$E(XY)=\sum_{x,y}xyP(X=x,Y=y)=\sum_{x,y}xyP(X=x)P(Y=y)$$
$$=(\sum_{x}xP(X=x))(\sum_{y}yP(Y=y))=E(X)E(Y).$$

(b) 举一个简单的例子（见图 A-18）. 假设(X, Y)等可能地在如下 4 组数中取值：

$$\{(-1,0),(0,1),(0,-1),(1,0)\}.$$

则 $E(XY)=0=E(X)E(Y)$，所以 X 和 Y 是不相关的. 但是 $P(X=-1,Y=1)=0 \neq P(X=-1)P(Y=1)$，因此 X 和 Y 不独立.

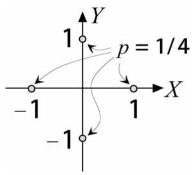


图 A-18　随机变量 X 和 Y 不相关也不独立

(c) 设 X 和 Y 是不相关的随机变量，则

$$\begin{aligned}\mathrm{var}(X+Y)&=E((X+Y-E(X+Y))^2)\\&=E(X^2+Y^2+2XY-E(X)^2-E(\mathrm{Y})^2-2E(X)E(Y))\\&=E(X^2)-E(X)^2+E(\mathrm{Y}^2)-E(\mathrm{Y})^2\\&=\mathrm{var}(X)+\mathrm{var}(Y).\end{aligned}$$

其中第三个等式成立的原因是 $E(XY)=E(X)E(Y)$.　♣

上述结论可以推广到非离散的情形. 对于一对随机变量(X, Y)，我们定义其联合累积概率分布为 $F_{X,Y}(x,y)$，满足如下条件：

$$F_{X,Y}(x,y)=P(X\leqslant x,Y\leqslant y),\ x,y\in\Re.$$

当这个函数的导数存在时，定义它为联合概率密度函数，记为 $f_{X,Y}(x,y)$. 即

$$F_{X,Y}(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}f_{X,Y}(u,v)\,\mathrm{d}u\,\mathrm{d}v.$$

联合概率密度函数的含义是

$$f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y=P(X\in(x,x+\mathrm{d}x)\text{且}Y\in(y,y+\mathrm{d}y)).$$

例如，假设

$$f_{X,Y}(x,y)=\frac{1}{\pi}1\{x^2+y^2\leqslant 1\},\ x,y\in\Re.$$

则称 (X,Y) 均匀且随机地分布在单位圆内.

理解随机变量的一个直观方法是，在平面上以 $\varepsilon>0$ 为间距画出网格，并用如下方法近似表达 (X,Y)：

$$(X,Y)=(m\varepsilon,n\varepsilon)\ w.p.\ f_{X,Y}(m\varepsilon,n\varepsilon)\varepsilon^2.$$

在这种近似下，

$$E(h(X,Y))=\sum_{m,n}h(m\varepsilon,n\varepsilon)f_{X,Y}(m\varepsilon,n\varepsilon)\varepsilon^2$$
$$\approx\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}h(x,y)\mathrm{f}_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

下面将其整理成为正式的定义.

定义 A.7 设 (X,Y) 是一对随机变量，函数 $h:\mathfrak{R}\to\mathfrak{R}$. 当联合概率密度函数存在时，有

$$E(h(X,Y)):=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}h(x,y)f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y. \qquad \Diamond$$

该定义保证了期望是线性的. 和前面的思路相仿，我们可以定义 X 和 Y 的协方差等概念.

定义 A.8 独立

如果

$$P(X\in A,Y\in B)=P(X\in A)P(Y\in B)$$

对于任意实数集 A 和 B 成立，则称随机变量 X 和 Y 是独立的. $\Diamond$

经过简单的推算可以得到，当联合概率密度函数存在时，两个随机变量独立的条件是当且仅当

$$f_{X,Y}(x,y)=f_X(x)f_Y(y),\ \forall x,y\in\mathfrak{R}.$$

对于给定的随机变量 X 和任意函数 $g:\mathfrak{R}\to\mathfrak{R}$， $g(X)$ 也是一个随机变量. 注意

$$g(X)\in A.\ \text{当且仅当} X\in g^{-1}(A):=\{x\in\mathfrak{R}\,|\,g(x)\in A\}.$$

当然，它们实际上是一回事.

下面是一条很有用的结论.

定理 A.9 独立的随机变量的函数是相互独立的

设 X 和 Y 是相互独立的随机变量，g 和 $h:\mathfrak{R}\to\mathfrak{R}$ 是任意的函数，则 $g(X)$ 与 $h(Y)$ 是两个相互独立的随机变量. □

证明

注意

$$
\begin{aligned}
P(g(X)\in A,h(Y)\in B) &= P(X\in g^{-1}(A),Y\in h^{-1}(B)) \\
&= P(X\in g^{-1}(A))P(Y\in h^{-1}(B)) \\
&= P(g(X)\in A)P(h(Y)\in B).
\end{aligned}
$$

♣

A.4.1 相互独立随机变量的最大值与最小值

人们经常需要考虑相互独立随机变量的最大值与最小值，基本的想法如下．设 X 和 Y 是相互独立的随机变量，并且 $V=\min\{X,Y\}$ 、$W=\max\{X,Y\}$ ，则

$$P(V>v)=P(X>v,Y>v)=P(X>v)P(Y>v).$$

同时，

$$P(W\leqslant w)=P(X\leqslant w,Y\leqslant w)=P(X\leqslant w)P(Y\leqslant w).$$

这些结论足以进行很多有用的计算．

例如，设 $X=Exp(\lambda)$ 、$Y=Exp(\mu)$ ，则

$$P(V>v)=P(X>v)P(Y>v)=\exp\{-\lambda v\}\exp\{-\mu v\}=\exp\{-(\lambda+\mu)v\}.$$

因此，2 个指数分布随机变量的最小值仍然服从指数分布，其率参数等于原来的率参数之和．

设 X 和 Y 是独立同分布的 $U[0,1]$ 变量，则

$$P(W\leqslant w)=P(X\leqslant w)P(Y\leqslant w)=w^2,\ w\in[0,1].$$

设 A 和 B 是两个事件，其中 $P(B)>0$．在已知 B 的情况下，定义 A 的条件概率 $P[A|B]$如下：

$$P[A\,|\,B]:=P(A\cdot B)/P(B).$$

$P[A|B]$表示的是，已知实验结果落在事件 B 定义的集合内，则该结果落在事件 A 定义集合的概率．举例来说，假设一个随机性实验有 1000 个等可能出现的结果．并假设事件 A 包含$|A|$个可能出现的结果，事件 B 包含$|B|$个可能出现的结果．如果已经知道结果落在集合 B 内，那么它可能是$|B|$个结果中的任意一个．此时，结果同时落在集合 A 中的概率就等于集合 B 中的结果同时属于集合 A 的比例，即

$$\frac{|A\cdot B|}{|B|}=\frac{|A\cdot B|/1000}{|B|/1000}=\frac{P(A\cdot B)}{P(B)}.$$

注意到，根据这个定义，当 A 和 B 相互独立时，我们有 $P[A\,|\,B]=P(A)$ ．这个结果非常直观．同时得到：

$$P(A\cdot B)=P[A\,|\,B]\cdot P(B).$$

A.4.2 独立随机变量之和

设 X 和 Y 是独立的随机变量，并且 $Z=X+Y$．我们希望借助 $f_X(x)$ 和 $f_Y(y)$ 来计算 $f_Z(z)$，思路是

$$P(Z\in(z,z+\mathrm{d}z))=\int_{-\infty}^{\infty}P(X\in(x,x+\mathrm{d}x),Y\in(z-x,z-x+\mathrm{d}z)).$$

因此，

$$f_Z(z)\mathrm{d}z=\int_{-\infty}^{\infty}f_X(x)f_Y(z-x)\mathrm{d}x\mathrm{d}z.$$

即

$$f_Z(z)=\int_{-\infty}^{\infty}f_X(x)f_Y(z-x)\mathrm{d}x=f_X*f_Y(z),$$

这里 $g*h$ 表示两个函数的卷积. 如果上过信号与系统这门课，你就会明白如何采用“平移与反褶”作图法来计算卷积.

A.5 随机向量

在很多情况下，我们感兴趣的是随机变量组合的性质.

定义 A.10 随机向量

随机向量 $\boldsymbol{X}=(X_1,\cdots,X_n)'$ 是以随机变量为元素构成的．它的性质可以用**联合累积分布函数**来表示：

$$F_{\boldsymbol{X}}(x_1,\cdots,x_n):=P(X_1\leqslant x_1,\cdots,X_n\leqslant x_n),x_i\in\Re,i=1,\cdots,n.$$

相应的联合概率密度函数 $f_{\boldsymbol{X}}(\boldsymbol{x})$ 满足：

$$F_{\boldsymbol{X}}(x_1,\cdots,x_n)=\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n}f_{\boldsymbol{X}}(u_1,\cdots,u_n)\,\mathrm{d}u_1\cdots\mathrm{d}u_n,$$

前提是这个函数存在．此时，

$$f_{\boldsymbol{X}}(\boldsymbol{x})\mathrm{d}x_1\cdots\mathrm{d}x_n=P(X_i\in[x_i,x_i+\mathrm{d}x_i],i=1,\cdots,n). \qquad \Diamond$$

综上所述，联合累积分布函数和联合概率密度函数描述了该随机向量的取值落在 $\Re^n$ 某个子集中的概率.

与两个随机变量的情形类似，当联合概率密度函数存在时，可以得到

$$E(h(\boldsymbol{X}))=\int\cdots\int h(x)f_X(u)\mathrm{d}u_1...\mathrm{d}u_n.$$

下面是一个我们会经常用到的定义.

定义 A.11 期望与协方差

设 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 是随机向量. 定义

$$E(\boldsymbol{X})=(E(X_1),\cdots,E(X_n))'$$
$$\sum\nolimits_X=E((\boldsymbol{X}-E(\boldsymbol{X}))(\boldsymbol{X}-E(\boldsymbol{X}))')$$
$$\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})=E((\boldsymbol{X}-E(\boldsymbol{X}))(\boldsymbol{Y}-E(\boldsymbol{Y}))').$$

当 $\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})=0$ 时，即对于所有的 i 和 j 都有 X_i 和 Y_j 不相关，则称 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 是**不相关的**. ◊

因此，随机向量的期望就是各分量期望组成的向量. 类似地，随机矩阵的期望就是各分量期望组成的矩阵. 随机向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的协方差就是各分量协方差组成的矩阵. 事实上，

$$\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})_{i,j}=E((X_i-E(X_i))(Y_j-E(Y_j)))=\operatorname{cov}(X_i,Y_j).$$

同时，$\sum_X=\operatorname{cov}(\boldsymbol{X},\boldsymbol{X})=:\operatorname{cov}(\boldsymbol{X})$.

作为简单的练习，请读者验证

$$\operatorname{cov}(A\boldsymbol{X}+\boldsymbol{a},B\boldsymbol{Y}+\boldsymbol{b})=A\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})B'.$$

正交与投影

在估计问题中，正交与投影的概念是很重要的.

设 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 是两个随机向量. 如果

$$E(\boldsymbol{X}\boldsymbol{Y}')=0,$$

则称 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 正交，记为 $\boldsymbol{X}\perp\boldsymbol{Y}$. 因此，$X$ 与 Y 正交当且仅当每一个 X_i 正交于每一个 Y_j.

注意到当 $E(\boldsymbol{X})=\boldsymbol{0}$ 时，$\boldsymbol{X}\perp\boldsymbol{Y}$ 当且仅当 $\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})=0$. 事实上，因为 $E(\boldsymbol{X})=\boldsymbol{0}$，所以

$$\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})=E(\boldsymbol{X}\boldsymbol{Y}')-E(\boldsymbol{X})E(\boldsymbol{Y})'=E(\boldsymbol{X}\boldsymbol{Y}'),$$

下面是一个很有用的结论（见图 A-19）.

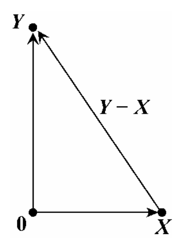


图 A-19 几何视角的正交

定理 A.12 正交

若 $\boldsymbol{X} \perp \boldsymbol{Y}$，则

$$E(\|\boldsymbol{Y}-\boldsymbol{X}\|^2) = E(\|\boldsymbol{X}\|^2) + E(\|\boldsymbol{Y}\|^2).$$

这个结论与勾股定理是等价的. □

证明

$$\begin{aligned} E(\|\boldsymbol{Y}-\boldsymbol{X}\|^2) &= E((\boldsymbol{Y}-\boldsymbol{X})'(\boldsymbol{Y}-\boldsymbol{X})) = E(\boldsymbol{Y}'\boldsymbol{Y}) - 2E(\boldsymbol{X}'\boldsymbol{Y}) + E(\boldsymbol{X}'\boldsymbol{X}) \\ &= E(\|\boldsymbol{Y}\|^2) - 2E(\boldsymbol{X}'\boldsymbol{Y}) + E((\boldsymbol{X})^2). \end{aligned}$$

如果 $\boldsymbol{X} \perp \boldsymbol{Y}$，则 $E(X_iY_j) = 0$ 对所有 i 和 j 均成立. 那么 $E(\boldsymbol{X}'\boldsymbol{Y}) = \sum_i E(X_iY_j) = 0$. 结论得证. ♣

A.6 随机变量的函数的概率密度

假如随机向量 $\boldsymbol{X}$ 是 $\Re^n$ 上的随机变量并具有概率密度函数 $f_{\boldsymbol{X}}(\boldsymbol{x})$，函数 $g: \Re^n \to \Re^n$ 是一个可导函数. 设 $\boldsymbol{Y} = g(\boldsymbol{X})$，该怎样计算 $f_{\boldsymbol{Y}}(\boldsymbol{y})$ 呢?

下面从线性函数的情况入手，然后解释一般的情况.

A.6.1 线性变换

设 X 的概率密度函数为 $f_X(x)$，$Y = aX + b(a > 0)$. 该怎样计算 $f_Y(y)$ 呢?

如图 A-20 所示,

$$\begin{aligned} P(Y \in (y, y+\mathrm{d}y)) &= P(aX + b \in (y, y+\mathrm{d}y)) \\ &= P(\mathrm{X} \in (a^{-1}(y-b), a^{-1}(y+\mathrm{d}y-b))). \end{aligned}$$

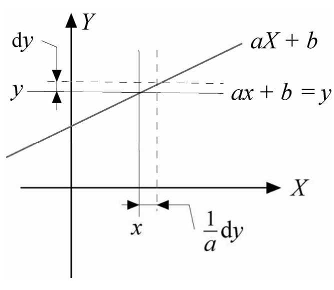


图 A-20 线性变换 $Y = aX+b$

因为 $P(Z \in (z, z+\mathrm{d}z)) = f_z(Z)\mathrm{d}z$，所以:

$$f_Y(y)\mathrm{d}y = f_X(a^{-1}(y-b)) \times a^{-1}\mathrm{d}y.$$

因此

$$f_Y(y) = \frac{1}{a} f_X(x)\text{，其中 } aX + b = y\ . \quad \text{（A.4）}$$

$a < 0$ 的情况与之类似．重复上述推理过程，可以得到

$$f_Y(y) = \frac{1}{|a|} f_X(x)\text{，其中 } aX + b = y\ .$$

那么对于二维的随机向量，情况又是如何呢？设 $\boldsymbol{X}$ 是一个随机向量，它在 $\Re^2$ 上取值，概率密度函数为 $f_{\boldsymbol{X}}(\boldsymbol{X})$．令

$$\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b},$$

这里 $A \in \Re^{2\times 2}$，$\boldsymbol{b} \in \Re^2$．

图 A-21 表明，在线性变换下，矩形 $[x_1, x_1 + \mathrm{d}x_1] \times [x_2, x_2 + \mathrm{d}x_2]$ 变换成为一个平行四边形，其面积为 $|A|\mathrm{d}x_1\mathrm{d}x_2$，$|A|$ 表示矩阵 A 行列式的绝对值．所以，$\boldsymbol{Y}$ 的取值落在面积 $|A|\mathrm{d}x_1\mathrm{d}x_2$ 中的概率为 $f_{\boldsymbol{X}}(\boldsymbol{x})\mathrm{d}x_1\mathrm{d}x_2$．同时，由于 $\boldsymbol{Y}$ 的取值落在某个区域内的概率与这个区域的面积成正比，Y 的取值落在面积 $|A|\mathrm{d}x_1\mathrm{d}x_2$ 中的概率也可以表示为 $f_{\boldsymbol{Y}}(\boldsymbol{y})|A|\mathrm{d}x_1\mathrm{d}x_2$，这里 $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{b}$．因此，

$$f_Y(y)|A|\mathrm{d}x_1\mathrm{d}x_2 = f_X(x)\mathrm{d}x_1\mathrm{d}x_2\text{，其中 } y = Ax + b\ .$$

所以

$$f_Y(y) = \frac{1}{|A|} f_X(x)\text{，这里 } y = Ax + b\ .$$

事实上，这个结论对 n 个随机变量也同样适用．

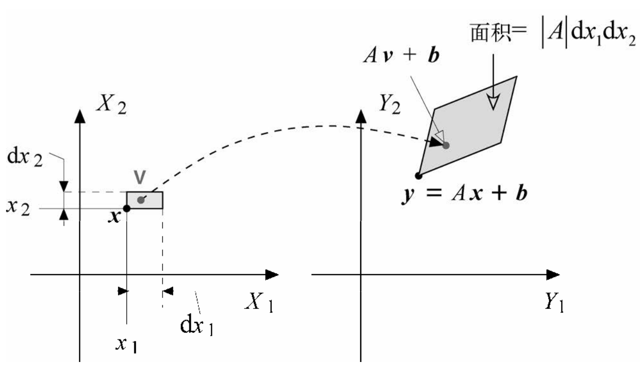


图 A-21 线性变换 $\boldsymbol{Y} = A\boldsymbol{X}+\boldsymbol{b}$

由于这个结论十分重要，将其总结为如下定理．

定理 A.13　线性变换中概率密度的变化

设 $\boldsymbol{Y}=A\boldsymbol{X}+\boldsymbol{b}$，其中 A 是一个 $n\times n$ 的非奇异矩阵. 那么

$$f_Y(\boldsymbol{y})=\frac{1}{|A|}f_X(\boldsymbol{x})\text{，其中 } A\boldsymbol{x}+\boldsymbol{b}=\boldsymbol{y}. \tag{A.5}$$

□

当 A 为奇异矩阵时，随机向量 $\boldsymbol{Y}=A\boldsymbol{X}+\boldsymbol{b}$ 只在小于 n 的维数中取值. 这种情况下，向量 $\boldsymbol{Y}$ 在 $\Re^n$ 中没有密度分布的概念. 作为一个简单的例子，设 X_1 和 X_2 相互独立且均匀分布在[0, 1]（见图 A-22）.

设

$$\boldsymbol{Y}=\begin{bmatrix}X_1\\X_1\end{bmatrix}=A\boldsymbol{X}\text{，其中 } A=\begin{bmatrix}1&0\\1&0\end{bmatrix}.$$

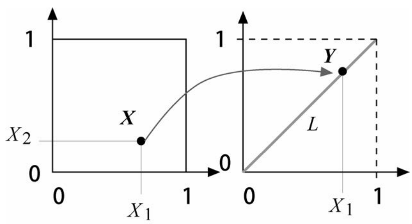


图 A-22　奇异变换 $\boldsymbol{Y}=(X_1, X_1)'$

那么 $\boldsymbol{Y}$ 在 $\Re^2$ 中是没有概率密度的. 假设 $\boldsymbol{Y}$ 在 $\Re^2$ 中有概率密度，则

$$L=\{\boldsymbol{y}|y_1=y_1\text{且}0\leqslant y_1\leqslant 1\},$$

$$P(\boldsymbol{Y}\in L)=\iint_L f_Y(\boldsymbol{y})\mathrm{d}y=0.$$

最后一个等式成立的原因是 L 在 $\Re^2$ 中的测度为 0，但是 $P(\boldsymbol{Y}\in L)=1$.

A.6.2　非线性变换

当 $Y=g(X)$ 且 $g(\cdot)$ 为非线性函数时，情况更为复杂. 首先来看一个例子.

例 1

设 $X=_D U[0,1]$ 且 $Y=X^2$，如图 A-23 所示. 对于 $0<\varepsilon\ll 1$，$Y\in[y,y+\varepsilon]$ 当且仅当 $X\in[x_1,x_1+\delta)$，这里

$$\delta=\frac{\varepsilon}{g'(x_1)}=\frac{\varepsilon}{2x_1}\text{，其中 } g(x_1)=x_1^2=y.$$

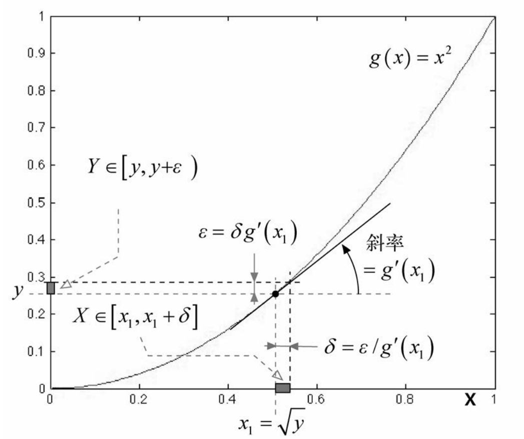


图 A-23 $X={}_DU[0,1]$时的变换 $Y=X^2$（另见彩插）

现在，

$$P(Y\in[y,y+\varepsilon))=f_Y(y)\varepsilon+o(\varepsilon)^{①}$$

而且

$$P(X\in[x_1,x_1+\delta))=f_X(x_1)\delta+o(\delta).$$

再加上 $o(\delta)=o(\varepsilon)$，我们得到：

$$f_Y(y)\varepsilon+o(\varepsilon)=f_X(x_1)\delta+o(\varepsilon)=\frac{1}{g'(x_1)}f_X(x_1)\varepsilon+o(\varepsilon).$$

所以

$$f_Y(y)=\frac{1}{g'(x_1)}f_X(x_1),\ g(x_1)=y.$$

在这个例子中，我们有 $g'(x_1)=2x_1=2\sqrt{y}$，且 $f_X(x_1)=1$．因此，

$$f_Y(y)=\frac{1}{2\sqrt{y'}}.$$

例 2

现在来看一个更为复杂的例子．设 $Y=g(X)=X^2$，X 取值范围为[−1, 1]，概率密度函数为

$$f_X(x)=\frac{3}{8}(1+x)^2,\ x\in[-1,1].$$

---

① $o(\varepsilon)$ 表示了这样一个关于 ε 的函数：$\frac{o(\varepsilon)}{\varepsilon}\to 0,\ \varepsilon\to 0$．

考虑一个 y 的取值 $y \in (0,1)$. 注意到现在有两个 x 的取值可以使 $g(x)=y$ ，分别是 $x_1=\sqrt{y}$ 和 $x_2=-\sqrt{y}$. 因此，

$$P(Y \in (y, y+\varepsilon)) = P(X \in (x_1, x_1+\delta_1)) + P(X \in (x_2-\delta_2, x_2)),$$

其中

$$\delta_1 = \frac{\varepsilon}{g'(x_1)}\ ,\quad \delta_2 = \frac{\varepsilon}{|g'(x_2)|}\,.$$

因此，

$$f_Y(y)\varepsilon + o(\varepsilon) = \frac{\varepsilon}{g'(x_1)} f_X(x_1) + \frac{\varepsilon}{\left|g'(x_2)\right|} f_X(x_2) + o(\varepsilon).$$

由此可以推出

$$f_Y(y) = \frac{1}{g'(x_1)} f_X(x_1) + \frac{1}{\left|g'(x_2)\right|} f_X(x_2)\,.$$

在这个例子中，我们有

$$f_Y(y) = \frac{1}{2\sqrt{y}}\frac{3}{8}(1+\sqrt{y})^2 + \frac{1}{2\sqrt{y}}\frac{3}{8}(1-\sqrt{y})^2 = \frac{3}{8}\frac{1+y}{\sqrt{y}}.$$

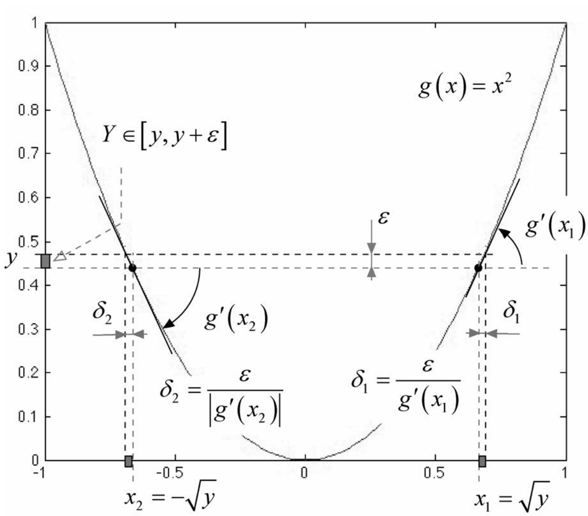


图 A-24　$X \in [-1,1]$ 时的变换 $Y = X^2$（另见彩插）

例 3

下一个例子针对一般的可导函数 $g(\cdot)$. 从例 2 可以看出，如果 $Y = g(X)$，那么

$$f_Y(y) = \sum_i \frac{1}{|g'(x_i)|} f_X(x_i), \tag{A.6}$$

并对所有使得 $g(x_i) = y$ 的 x_i 进行求和.

例 4

多维的情况如何呢？和前面一样，重要的思路是：从局部来看，$\boldsymbol{x}$ 到 $\boldsymbol{y}$ 的变换是线性的. 注意到

$$g_i(\boldsymbol{x} + \mathrm{d}\boldsymbol{x}) \approx g_i(\boldsymbol{x}) + \sum_j \frac{\partial}{\partial x_j} g_i(\boldsymbol{x}) \mathrm{d}x_j \approx g(\boldsymbol{x}) + J(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

这里矩阵 $J(\boldsymbol{x})$ 的[i，j]分量为

$$J_{i,j}(\boldsymbol{x}) = \frac{\partial}{\partial x_j} g_i(\boldsymbol{x}).$$

这是函数 $g: \Re^n \to \Re^n$ 的雅可比矩阵. 所以，从局部来看，这个变换可以近似为 $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b}$，其中 $\boldsymbol{b} = g(\boldsymbol{x})$，$A = J(\boldsymbol{x})$. 那么，在使得 $g(\boldsymbol{x}) = \boldsymbol{y}$ 的 $\boldsymbol{x}$ 附近，f_X 可以作近似于线性的变换：它由 $J(\boldsymbol{x})$ 的行列式进行了拉伸.

总结上述示例，我们得到如下定理.

定理 A.14　随机变量函数的密度

设 $\boldsymbol{X}$ 为 $\Re^n$ 上的随机向量并具有概率密度函数 f_X，且 $\boldsymbol{Y} = g(\boldsymbol{X})$，函数 $g: \Re^n \to \Re^n$ 是可导函数，那么

$$f_Y(\boldsymbol{y}) = \sum_i \frac{1}{|J(\boldsymbol{x}_i)|} f_X(\boldsymbol{x}_i).$$

这里对所有使得 $\boldsymbol{g}(\boldsymbol{x}_i) = \boldsymbol{y}$ 的 $\boldsymbol{x}_i$ 求和. $|J(\boldsymbol{x}_i)|$ 表示雅可比矩阵行列式在 $\boldsymbol{x}_i$ 点的绝对值. □

我们举一个例子来论述这一结论. 设 $\boldsymbol{X} = (X_1, X_2)$，其中 X_i 是独立同均匀分布的 $U[0,1]$ 随机变量. 考虑如下变换：

$$Y_1 = X_1^2 + X_2^2,\ Y_2 = 2X_1X_2, .$$

则

$$J(\boldsymbol{x}) = \begin{bmatrix} 2x_1 & 2x_2 \\ 2x_2 & 2x_1 \end{bmatrix}.$$

那么

$$|J(\boldsymbol{x})| = 4|x_1^2 - x_2^2|.$$

对于 $\boldsymbol{y}$ 的每一组取值，$\boldsymbol{x}$ 都有两组取值与之对应，分别是

$$x_1=\frac{1}{2}\left[\sqrt{y_1+y_2}+\sqrt{y_1-y_2}\right] \text{和} x_2=\frac{1}{2}\left[\sqrt{y_1+y_2}-\sqrt{y_1-y_2}\right]$$

以及

$$x_1=\frac{1}{2}\left[\sqrt{y_1+y_2}-\sqrt{y_1-y_2}\right] \text{和} x_2=\frac{1}{2}\left[\sqrt{y_1+y_2}+\sqrt{y_1-y_2}\right].$$

由这些取值得到

$$|J(\boldsymbol{x})|=\sqrt{y_1^2-y_2^2}.$$

所以，对于任意的 $\boldsymbol{y}$ 值，我们有：

$$f_Y(\boldsymbol{y})=\frac{2}{\sqrt{y_1^2-y_2^2}}.$$

A.7 参考资料

要完全想掌握概率论，好奇心、直觉和耐心都是必备的素质，当然也少不了好书的帮助．对我个人而言，Pitman 的 *Probability* 大有裨益．David Aldous 的主页提供了很多充满智慧、启发灵感的概率论评述．教材 *Introduction to Probability*、*Probability and Random Processes* 和 *Probability and Measure* , *Third Edition* 都很有参考价值．Wong 和 Hajek 的 *Stochastic Processes in Engineering Systems* 一书对本书话题提供了更为深入的讨论．强烈推荐 Gallager 的新书 *Stochastic Processes: Theory for Applications*，它能够很好地补充本书内容.

互联网是搜集各种资料的强大工具，包括概率论．我经常开玩笑说："不需要记笔记，这些知识都在网上呢."

A.8 练习题

1. 判断下列论述是否正确.

(a) 没有交集的事件是相互独立的.

(b) 随机变量之和的方差总是它们的方差之和.

(c) 随机变量之和的期望总是它们的期望之和.

2. 举出满足下列条件的事件 A、B、C：

$P[A|C]<P(A)$，$P[A|B]>P(A)$ 并且 $P[B|A]>P(B)$.

3. 抛掷两枚均匀的骰子，记事件 A 为"两枚骰子点数之和小于等于 8"，记事件

B 为“第一枚骰子的点数大于等于 3”.

(a) 这里的概率空间 $(\Omega, \mathcal{F}, P)$ 是什么?

(b) 计算 $P[A|B]$和 $P[B \mid A]$.

4. 一个人不停地抛掷一枚均匀硬币.

(a) 在开始抛掷的前 1000 次中，正面朝上的次数是偶数的概率是多少?

(b) 在开始抛掷的前 4 次中，正面朝上次数多于背面朝上的概率是多少?

5. 一个人不停地抛掷一枚均匀硬币.

试证明：抛出正面的次数总是多于抛出背面的概率为 0.

6. 设 X 和 Y 是率系数为 1 的独立指数变量. 求 $Z = X+Y$ 的概率密度函数.

7. 从洗匀的 52 张扑克牌中随机抽取. 假设抽取了 4 张牌,数值都在 2 到 10 之间,如方块 2、红桃 10、草花 6 以及黑桃 2. 试编写 Matlab 程序，计算黑色扑克牌上数值之和恰好是红色扑克牌上数值之和 2 倍的概率.

8. 考虑[0, 1]上的均匀分布. 在[0, 1]区间上找到 3 个子区间 A、B、C,使得 $P[A|B] > P(A)$且 $P[C|B] < P(C)$.

9. 设 $X =_D G(p)$，即 X 是关于参数 p 的几何分布. 计算 $E(X^3)$.

10. 设 X 和 Y 为独立同分布的 $U_D[0, 1]$ 变量. 计算 $E(\max\{X,Y\} - \min\{X,Y\})$.

11. 设 $X =_D P(\lambda)$，即 X 是关于参数 λ 的泊松分布. 求 P（X 为偶数）.

12. 考虑 Ω=[0, 1] 的均匀分布随机变量. 设 $X(\omega) = 1\{a < \omega < b\}$ 且 $Y = 1\{c < \omega < d\}$，其中 $0 < a < b < 1$，$0 < c < d < 1$. 假设 X 和 Y 是不相关的，那么它们是否一定独立?

13. 设 X 和 Y 为独立同分布的 $U[-1, 1]$ 变量，并定义 $Z = XY$. X 和 Z 是不相关的吗? 它们相互独立吗?

14. 设 $X =_D U[-1,3]$，$Y = X^3$. 计算 $f_Y(\cdot)$.

15. 有一根 1 米长的棍子，沿着棍子随机且独立地选取点 X 和 Y，并从这两点处截断. 试求可以用截得的三段拼成一个三角形的概率?

16. 有两个朋友偶尔去酒吧,他们去的时间都均匀分布在下午 5:00~6:00 且相互独立. 每个人到达后，都会等待另一位朋友 10 分钟. 请问他们能相遇的概率是多大?

17. 选取 $V \geqslant 0$ 使得 $V^2 =_D Exp(2)$，并选取独立于 V 的 $\theta =_D U[0, 2\pi]$. 设 $X = V\cos(\theta)$，$Y = V\sin(\theta)$. 计算 $f_{X,Y}(x, y)$.

18. 设 Z 和 $1/Z$ 是具有相同概率分布的随机变量,并且存在 $E(|Z|)$. 证明 $E(|Z|) \geqslant 1$.

19. 设 $\{X_n, n \geqslant 1\}$ 是期望为 0、方差为 1 的独立同分布随机变量. 设 $Y_n = (X_1 + \cdots + X_n)/n$.

(a) 计算 $\mathrm{var}(Y_n)$.

(b) 证明 $P(|Y_n| \geqslant \varepsilon) \to 0$，$n \to \infty$ 对于所有的 $\varepsilon > 0$ 都成立.

20. 设 X 和 Y 为独立同分布的 $U[0,1]$ 变量，$Z = A(X,Y)^T$,其中 A 是已知的 2×2 矩

阵．求 Z 的概率密度函数．

21. 设 $X =_D U[1,7]$，$Y = \ln(X) + 3\sqrt{X}$．证明 $E(Y) \leqslant 7.4$．

22. 在 $[0, 1]^2$ 上独立且均匀地选取点 X 和 Y．试计算 $E(\|X - Y\|^2)$．

23. 设(X, Y)是在以$(-1, 0)$，$(1, 0)$，$(0, 1)$为顶点的三角形内随机选取的点．计算 $\mathrm{cov}(X, Y)$．

24. 设随机变量 X 的期望为 1，方差为 0.5．证明

$$E(2X + 3X^2 + X^4) \geqslant 8.5.$$

25. 设 X、Y、Z 独立同分布，均匀分布于$\{-1, 1\}$（即，等可能地取值−1 或 1）．设 $V_1 = XY$、$V_2 = YZ$、$V_3 = XZ$．

(a) $\{V_1, V_2, V_3\}$ 是两两独立的吗？请给出证明．

(b) $\{V_1, V_2, V_3\}$ 是相互独立的吗？请给出证明．

26. 设事件 A 和 B 的概率为 $P(A) = 3/4$，$P(B) = 1/3$．证明 $\dfrac{1}{12} \leqslant P(A \cap B) \leqslant \dfrac{1}{3}$，并举例说明这里的上界和下界都无法再作改进．找出 $P(A \cup B)$ 的取值范围．

27. 一个例子，使得事件 A、B 和 C 两两独立，但并不是相互独立．

28. 发电站用 N 台发电机为城市供电．每一台发电机出故障的概率都是 p，彼此独立．如果少于 $k(0 \leqslant k \leqslant N)$ 台发电机供电，该城市将会断电．试求断电的概率？写一段 Matlab 程序，计算对于任意的 N、k 和 p，这个概率值是多大．设 $N = 100$，$p = 0.1$，画出 k 从 0 到 N 变化时的断电概率曲线．

29. 图 A-25 展示了一个系统的可靠性模型．图中的边代表系统中的元素．编号为 i 的边有 p_i 的概率正常工作，有 $1-p_i$ 的概率不工作，每条边与其他的边独立．当节点 S 和 T 被连接起来时，这个系统能够正常工作．也就是说，这个系统由两个可以相互替代的子系统构成．每一个子系统都由一些元素构成．

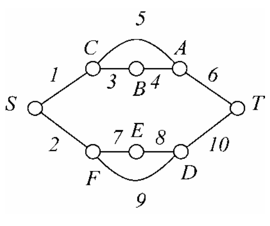


图 A-25　系统的可靠性模型

(a) 计算这个系统能正常工作的概率．

(b) 假设这个可靠性模型是 n 层的二叉树，每一条边失效的概率是 p．从根节点到某一个叶子节点存在工作通路的概率是多大？

30. 设 X 和 Y 为独立同分布的 $U[0,1]$ 变量，且 $V = 3X + Y$，$W = X + Y$．计算

$f_{V,W}(v,w)$.

31. 设随机变量 X, Y 为独立同分布的 $U[0,1]$ 变量，且 $V = X+Y$，$W = X-Y$.

(a) 证明 V 和 W 是不相关的.

(b) V 和 W 是独立的吗？证明你的结论.

32. 图 A-26 展示的 RC 电路被用作简单计时器. 起初，电容通过电源充电 5V. 在 $t=0$ 时拨动开关，电容开始通过电阻放电. 有一个外部电路检测电压 $V(t)$降低到小于 1V 需要的时间 τ.

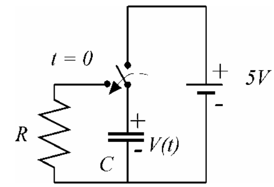


图 A-26　用作简单计时器的电路. 外部电路检测电压 $V(t)$降低到小于 1V 的时间

(a) 计算出 τ 关于 R 和 C 的表达式.

(b) 现在假设 R 和 C 是相互独立的随机变量，分别均匀分布在 $[R_0(1-\varepsilon), R_0(1+\varepsilon)]$ 和 $[C_0(1-\varepsilon), C_0(1+\varepsilon)]$ 上. 计算 τ 的方差.

(c) 设 τ_0 是 $R = R_0$ 且 $C = C_0$ 时 τ 的取值. 对于一个很小的值 δ，找到 $|\tau-\tau_0| > \delta\tau_0$ 的概率最大值.

33. 甲和乙一起玩硬币配对的游戏. 在这个游戏中，他们分别决定展示自己手中硬币的哪一面. 当两枚硬币不同时，甲获胜；否则乙获胜.

(a) 假设甲选择正面的概率 $P_A \in [0,1]$. 计算乙展示正面的概率 P_B 取多少才能使他获胜的概率最大.

(b) 依据你计算出的结果，分别找到对甲和乙来说 P_A 和 P_B 的最佳取值. 请说明在这样的取值下，双方不能进一步通过改变 P_A 和 P_B 来提高获胜的概率. 具有这种性质的方案被称为纳什均衡.



图 A-27　甲和乙玩硬币配对游戏

34. 你在玩某一种概率游戏，每一步获胜的概率都是 p. 你的起始资金是 1 美元. 每一步，如果获胜就可以得到 1 美元，失败就输掉 1 美元. 请计算在破产前能够积累到 4 美元的概率.

35. 你在抽屉里发现了 2 枚旧电池，它们能产生的电压值为 X 和 Y．设 X 和 Y 为独立同分布的 $U[0,1.5]$ 变量．

(a) 如果把 2 枚电池串联，能产生超过 2V 电压的概率是多少？

(b) 至少有 1 枚电池的电压值大于 1V 的概率是多少？

(c) 2 枚电池的电压值都大于 1V 的概率是多少？

(d) 你在抽屉里找到了更多类似的电池．逐一检测它们，直到找到 1 枚电池的电压值超过 1.2V．你需要检测的电池个数期望值是多少？

(e) 逐一检测这些电池，直到连续找到 2 个电压大于 1.2V 的电池．你需要检测的电池个数期望值是多少？

(f) 逐一检测这些电池并且记录其电压值，直到其中 2 个的电压值之和超过 2.5V．你需要检测的电池个数期望值是多少？



图 A-28　电池

36. 你想要卖掉旧手机．朋友甲和乙都对此感兴趣．你知道他们对手机的估价分别为 X 和 Y，并且 X 和 Y 均匀同分布于 $U[50,150]$．你提议进行如下的拍卖：你标价为 R，如果甲出价 A、乙出价 B，那么手机归出价较高者所有（前提是这个价位高于 R）．这位竞价者以标价和另一位竞价者出价中的较大值作为成交价．也就是说，如果 $A < R < B$，那么乙可以买到手机，付款 R．如果 $R<A<B$，那么乙可以买到手机，付款 A．

(a) 如果 $A = X$，$B = Y$，那么售价的期望值是多少？

(b) 找到 R 值，使期望得到的付款最多．

(c) 我们定义，当甲得到手机且付款为 P 时其盈余是 $X–P$；没有得到手机时盈余是 0．乙的盈余也同样加以定义．请证明甲出价 $A = X$ 时盈余的期望值最大；对乙也有类似的结论．我们称这一场拍卖是激励相容的，同时也是收益最大化的．

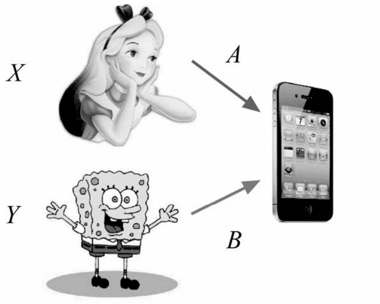


图 A-29　甲和乙对手机的估价分别为 X 和 Y，并分别出价 A 和 B

37. 方阵 S 的迹 $\mathrm{tr}(S)$定义为它的对角线元素之和．设 A 是一个 $m\times n$ 矩阵，B 是一个 $n\times m$ 矩阵．证明 $\mathrm{tr}(AB)=\mathrm{tr}(BA)$．

38. 设 Σ 是某个随机向量 $\boldsymbol{X}$ 的协方差．证明对于所有的实向量 $\boldsymbol{a}$ 都有 $\boldsymbol{a}'\Sigma\boldsymbol{a}\geqslant 0$．

39. 你希望给自己的房子购置太阳能电池板．折算太阳能电池板有效期内的花费以后，已知当电池板能传送的最大功率为 K 时，每单位时间需要的费用为 αK．假设太阳能电池板传送的实际功率 Z 服从均匀分布 $U[0,K]$．

所需电能 X 的分布为 $U[0,A]$，并且假设这和太阳能电池板传输的功率是独立的．如果太阳能电池板的最大功率是 K，那么单位时间内的花费是

$$\alpha K+\beta\max\{0,X-Z\},$$

其中后一项是需要从电网购买的电能．请计算所购买太阳能电池板的最大功率 K，使得单位时间的费用期望值最小．



图 A-30　应该为房子购置多大的太阳能电池板

40. 现在需要你设计一个散列表来存储与复原一个网络的数据．要存储元素的名称长度都是 n 比特，而你希望把它们存储在 m 比特的空间中．假设这些元素是均匀同分布的，随机分布在 $\{1,\cdots,2^m\}$．一共有 K 个元素需要存储．构造这个散列表的花费是 $\alpha 2^m$，一次冲突的花费是 β．当 2 个或 2 个以上元素占据散列表中相同位置时会发生冲突．试确定 m 的值，使得构造散列表及冲突带来的花费最少．

附录B

线性代数基本知识

> **主题**：矩阵及向量的操作，行列式，求逆，特征分解，若当标准型，奇异值分解

B.1 矩阵符号

处理多个实数值数据的时候，常常通过矩阵和向量来进行线性运算.

对于 $m \in \{1, 2, \cdots\}$，在 $\Re^m$ 中的每个列向量 $\boldsymbol{v}$ 都由 m 个实数纵向排列组成. 这 m 个实数叫作向量的元素. 下面是一些 $\Re^3$ 中列向量的例子：

$$\boldsymbol{v} = \begin{bmatrix} 2 \\ -1.5 \\ \sqrt{3} \end{bmatrix}, \quad \boldsymbol{w} = \begin{bmatrix} -0.2 \\ 3.1 \\ 0 \end{bmatrix}, \quad \boldsymbol{0} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

我们将向量 $\boldsymbol{x} \in \Re^m$ 的组成元素记为 $(x_1, x_2, \cdots, x_m)$. 行向量的定义与列向量类似，不过它们由组成元素的横向排列构成.

一般来说，对于给定的整数 m 和 n，矩阵 $A \in \Re^{m \times n}$ 是由实数按 m 行 n 列的矩形排列方式组成. 在矩阵 i 行 j 列的元素记为 $A(i,j)$. 当 $m=n$ 时，也称该矩阵为方阵.

同时定义单位矩阵 $\boldsymbol{I} \in \Re^{n \times n}$，$\boldsymbol{I}$ 的元素 $\boldsymbol{I}(i,j)=1\{i=j\}$. 也就是说，其对角线上元素等于 1，非对角线元素等于 0（n 的值通常可以通过上下文得到）.

根据定义，$\boldsymbol{v} \in \Re^{1 \times m}$ 是一个行向量，而 $\boldsymbol{w} \in \Re^{m \times 1}$ 是一个列向量.

B.2 矩阵运算

本节将回顾矩阵和向量的一些关键运算.

下面介绍矩阵的基本运算. 对于矩阵 $A, C, D \in \Re^{m \times n}$，$B \in \Re^{m \times k}$ 以及实数 $\alpha, \beta \in \Re$，有如下定义.

C.1

- 转置：将矩阵 A 变换成 A'，使得 $A'(i, j)=A(j, i)$.
- 相乘：AB 的定义为 $(AB)(i,j)=\sum_k A(i,k)B(k,j)$.
- 线性组合：$\alpha C+\beta D$ 的定义为 $(\alpha C+\beta D)(i,j)=\alpha C(i,j)+\beta D(i,j)$.

例如，

$$\begin{bmatrix}2 & 1 & -1\\ 3 & -1 & -2\end{bmatrix}'=\begin{bmatrix}2 & 3\\ 1 & -1\\ -1 & -2\end{bmatrix},\quad \begin{bmatrix}2 & 1 & -1\\ 3 & -1 & -2\end{bmatrix}\begin{bmatrix}1 & 5\\ -2 & 3\\ 0 & 4\end{bmatrix}=\begin{bmatrix}0 & 9\\ 5 & 4\end{bmatrix}$$

以及

$$2\begin{bmatrix}2 & 1 & -1\\ 3 & -1 & -2\end{bmatrix}-\begin{bmatrix}4 & 3 & -6\\ -2 & -4 & 3\end{bmatrix}=\begin{bmatrix}0 & -1 & 4\\ 8 & 2 & -7\end{bmatrix}.$$

从上述定义可以看到，矩阵的运算继承了实数运算的性质．例如，

$$(A+B)C=AC+BC \text{ 以及 } A(BC)=(AB)C.$$

同时，我们可以验证

$$(AB)'=B'A'.$$

如果 $\boldsymbol{x},\boldsymbol{y}\in\Re^{m\times 1}$，则 $\boldsymbol{x}'\boldsymbol{y}$ 称为 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的标量积．

如果 $A=[\boldsymbol{v}_1|\ldots|\boldsymbol{v}_n]$ 由列向量 $\boldsymbol{v}_j$ 组成，并且 $\boldsymbol{a}$ 也是一个列向量，则它们的乘积为

$$A\boldsymbol{a}=\sum_j a_j\boldsymbol{v}_j.$$

此外，如果矩阵 A 是由行向量 $\boldsymbol{r}_1, \cdots, \boldsymbol{r}_n$ 组成的，同时 $\boldsymbol{c}$ 是一个行向量，则

$$\boldsymbol{c}A=\sum_j c_j\boldsymbol{r}_j.$$

B.3 行列式

本节定义了方阵的行列式．这个概念对于逆矩阵的定义以及矩阵特征值的计算十分重要．

我们可以以递归的方式定义方阵的行列式．1×1 矩阵的行列式就等于其元素的值．假设我们已经定义好了 $n\times n$ 矩阵的行列式．设 A 是一个 $(n+1)\times(n+1)$ 的矩阵．对于 A 中第一列的每一个元素 $A(i, 1)$，定义 $B(i, 1)$是把 A 的第 1 列和第 i 行划掉以后得到的矩阵．然后，将 A 的行列式 $\det(A)$定义为：

$$\det(A)=\sum_i A(i,1)(-1)^{i+1}\det(B_{i,1})=:\sum_i A(i,1)C(i,1),$$

其中 $C(i,1)=(-1)^{i+1}\det(B_{i,1})$. 例如，

$$\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = a(-1)^{1+1}d + c(-1)^{2+1}b = ad - bc.$$

再举一个例子，可以通过归纳法得到

$$\det(\boldsymbol{I}) = 1.$$

另一个例子是，

$$\det\left(\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & k \end{bmatrix}\right) = a(ek - hf) - d(bk - hc) + g(bf - ec).$$

在最后这个例子里，注意行列式是一系列乘积的和，而每项乘积都由矩阵中的 3 个元素相乘得到，如 aek、ahf、dbk 等；在每一项乘积中，都没有任何两个元素在同一行或同一列．也就是说，每一个乘积项都具有 $A(1, j_1)A(2, j_2)A(3, j_3)$的形式，其中 j_1、j_2 和 j_3 均不相同．例如，$cdh=A(1, 3)A(2, 1)A(3, 2)$．在求和的过程中，每一个这样的乘积项都乘了+1 或−1．其中，如果将（j_1, j_2, j_3）变换成升序序列所需要的最少相邻项交换次数是偶数，则在 $A(1, j_1)A(2, j_2)A(3, j_3)$乘积项前乘+1；否则乘以−1．例如，$cdh=A(1, 3)A(2, 1)A(3, 2)$要乘以+1，因为可以通过 2 次相邻项的交换将 312 变成按升序排列：312→132→123.

一般来说，对于矩阵 $A \in \Re^{n\times n}$，我们有

$$\det(A) = \sum_p (-1)^{n(p)} A(1, p_1)A(2, p_2)\cdots A(n, p_n), \tag{B.1}$$

其中要对于所有 $12\cdots n$ 的排列 p 求和，$n(p)$表示为了将 p 变成升序所需要的相邻项的交换次数的最小值.

行列式还有另一种很有用的表示形式．对任意的(i, j)，设 $B_{i,j}$ 表示将矩阵 A 的第 i 行和第 j 列划掉得到的矩阵，设 $C(i, j) = (-1)^{i+j}\det(B_{i,j})$．这里的 $C(i, j)$ 称为元素 $A(i, j)$ 的余子式．则

$$\det(A) = \sum_i A(i, j)C(i, j), \forall j \text{，同时} \det(A) = \sum_j A(i, j)C(i, j), \forall i. \tag{B.2}$$

读者可以直接通过前面 2×2 和 3×3 的例子来验证这个表达式．通过仔细思考，我们会发现这个定义式是可以由式（B.1）推出来的．由这些表达式还可以推出

$$\det(A) = \det(A').$$

如果一个矩阵有两个相同的行或列，那么它的行列式等于 0．这对于所有的 2×2 矩阵都是成立的．假设这对于所有的 $n\times n$ 矩阵都成立．设 A 是一个 $(n+1)\times(n+1)$ 的矩阵，并设它的第 i 行和第 k 行是相同的．现在选一个不是 i 也不是 k 的数 $s\,(i \ne s \ne k)$．由式（B.2）可得

$$\det(A)=\sum_j A(s,j)C(s,j).$$

注意到$C(s,j)=(-1)^{s+j}\det(B_{s,j})$. 但是$B_{s,j}$有两个相同的行(原矩阵的第$i$行和第$k$行),所以$\det(B_{s,j})=0$. 因此,$\det(A)=0$. 上述结论对于所有的$n\geqslant 2$都成立. 有相同列的情况同样适用.

现在我们考虑由余子式$C(i,j)$组成的矩阵C. 我们有

$$AC'=\det(A)\boldsymbol{I}. \tag{B.3}$$

为了证明式(B.3)成立,首先考虑AC'的一个对角线元素. 我们有

$$(AC')(i,i)=\sum_j A(i,j)C'(j,i)=\sum_j A(i,j)C(i,j)=\det(A).$$

其中最后一个等式可以由式(B.2)得到. 下一步,我们考虑一个非对角线上的元素(i,k),这里$i\neq k$:

$$(AC')(i,k)=\sum_j A(i,j)C'(j,k)=\sum_j A(i,j)C(k,j).$$

现在,定义矩阵D为将矩阵A的第k行替换成A的第i行所得到的矩阵,那么

$$\det(D)=\sum_j D(i,j)E(i,j)=\sum_j A(i,j)E(i,j).$$

这里的$E(i,j)$表示$D(i,j)$的余子式. 那么$E(i,j)=(-1)^{i+j}\det(F_{i,j})$,其中$F_{i,j}$是划掉矩阵$D$中第$i$行和第$j$列得到的矩阵. 同时,$C(k,j)=(-1)^{k+j}\det(B_{k,j})$,其中$B_{k,j}$是划掉矩阵$A$中第$k$行和第$j$列得到的矩阵. 由$D$的定义可以看出$F_{i,j}=B_{k,j}$. 所以,对任意的$j$有$C(k,j)=(-1)^{k+i}E(i,j)$. 由此我们得到

$$(AC')(i,k)=(-1)^{i+k}\det(D).$$

前面已经说明$\det(D)=0$(因为D有两个相同的行),这就证明了式(B.3).

用同样的步骤可以得到

$$(C'A)=\det(A)\boldsymbol{I}. \tag{B.4}$$

行列式有一个非常重要的性质:如果$A,B\in\Re^{n\times n}$,则

$$\det(AB)=\det(A)\det(B). \tag{B.5}$$

这里我们不再证明这个式子. 读者可以参考其他的线性代数书籍.

B.4 逆矩阵

我们的核心结论是:方阵可逆当且仅当它的行列式非零. 此时,矩阵的列向量彼此之间是线性无关的.

设 $A \in \Re^{n\times n}$ 是一个方阵．如果存在一个矩阵 B 使得 $AB=\boldsymbol{I}$ 且 $BA=\boldsymbol{I}$ [①]，则称 B 是 A 的逆矩阵．

注意到，如果 A 的逆矩阵是 B，则由式（B.5）可知，$\det(A)\det(B) = \det(\boldsymbol{I})=1$，所以 $\det(A) \neq 0$.

反之，如果 $\det(A) \neq 0$，则它的逆矩阵可以表示为

$$A^{-1} = \frac{1}{\det(A)}C',$$

这里的矩阵 C 是由余子式组成的．由式（B.3）和式（B.4）可知

$$A\left(\frac{1}{\det(A)}C'\right) = \boldsymbol{I} \text{ 且 } \left(\frac{1}{\det(A)}C'\right)A = \boldsymbol{I}.$$

A 的逆矩阵是唯一的．事实上，如果 B 和 C 均为 A 的逆矩阵，则

$$C = (BA)\,C = B(AC) = B.$$

第一个等号成立是因为 $BA = \boldsymbol{I}$，第三个等号成立是因为 $AC = \boldsymbol{I}$.

我们还可以得到以下的结论，如果矩阵 A 和 B 都有逆矩阵存在，则矩阵 AB 也存在逆矩阵：

$$(AB)^{-1} = B^{-1}A^{-1}.$$

可以验证，$ABB^{-1}A^{-1} = AA^{-1} = \boldsymbol{I}$.

不仅如此，还有

$$(A')^{-1} = (A^{-1})'.$$

原因是 $A'(A^{-1})' = (A^{-1}A)^{-1} = \boldsymbol{I}$.

下面证明矩阵 A 存在逆矩阵当且仅当它的列向量 v_i 是线性无关的．根据线性无关的定义，这意味着不存在全为非零的系数能使得线性组合 $\sum_i a_i v_i$ 等于 0．也就是说，当且仅当 $\boldsymbol{a}=0$ 时 $A\boldsymbol{a}=0$．这一点不难看到：假设 A 存在逆矩阵且 $A\boldsymbol{a}=0$，那么

$$\boldsymbol{0} = A^{-1}A\boldsymbol{a} = \boldsymbol{a}.$$

所以 A 的列向量是线性无关的．为了证明逆命题，假设 A 的列向量是线性无关的．我们看到，对于所有的 $\boldsymbol{c} \in \Re^n$，方程

$$A\boldsymbol{x} = \boldsymbol{c}$$

至多有一个解 $\boldsymbol{x}$．事实上，如果上式有 2 个不同的解 $\boldsymbol{x}$ 和 $\boldsymbol{y}$，则 $A(\boldsymbol{x}-\boldsymbol{y})=0$，因为 A 的列向量线性无关，所以 $\boldsymbol{x}-\boldsymbol{y}=0$.

① 事实上，可以证明如果 $AB=\boldsymbol{I}$，则 $BA=\boldsymbol{I}$.

不仅如此，这个方程一定有一个解，我们稍后会论证这一点．因此，如果 A 的列向量线性无关，则 $A\boldsymbol{x}=\boldsymbol{c}$ 恰好有一个解．设这个解为 $g(\boldsymbol{c})$．方程 $g(\boldsymbol{c})$是 $\boldsymbol{c}$ 的线性函数，因为如果 $A\boldsymbol{x}_1=\boldsymbol{c}_1$ 且 $A\boldsymbol{x}_2=\boldsymbol{c}_2$，则 $A(\boldsymbol{x}_1+\boldsymbol{x}_2)=\boldsymbol{c}_1+\boldsymbol{c}_2$．所以，这个表达式可以写成 $g(\boldsymbol{c})=B\boldsymbol{c}$，其中 B 是一个矩阵①．也就是说，如果 $A\boldsymbol{x}=\boldsymbol{c}$，那么 $\boldsymbol{x}=B\boldsymbol{c}$．即 $BA\boldsymbol{x}=\boldsymbol{x}$．同样，如果 $\boldsymbol{x}=B\boldsymbol{c}$，那么 $A\boldsymbol{x}=\boldsymbol{c}$，所以 $AB=\boldsymbol{x}$．因此 $B=A^{-1}$．

现在来证明如果 A 的列向量线性无关，则 $A\boldsymbol{x}=\boldsymbol{c}$ 有一个解．证明的方法是设计一个具体的流程．下面通过一个例子来说明：

$$\begin{bmatrix}0 & 2 & 4\\ 3 & 6 & 9\\ 5 & 8 & 7\end{bmatrix}\boldsymbol{x}=\begin{bmatrix}14\\ 12\\ 18\end{bmatrix},$$

第一步推出

$$\begin{bmatrix}3 & 6 & 9\\ 0 & 2 & 4\\ 5 & 8 & 7\end{bmatrix}\boldsymbol{x}=\begin{bmatrix}12\\ 14\\ 18\end{bmatrix},$$

第二步推出

$$\begin{bmatrix}3 & 6 & 9\\ 0 & 2 & 4\\ 0 & -2 & -8\end{bmatrix}\boldsymbol{x}=\begin{bmatrix}12\\ 14\\ -2\end{bmatrix},$$

第三步推出

$$\begin{bmatrix}3 & 6 & 9\\ 0 & 2 & 4\\ 0 & 0 & -4\end{bmatrix}\boldsymbol{x}=\begin{bmatrix}12\\ 14\\ 12\end{bmatrix}.$$

这种算法通过行交换以及行的线性变换来实现对方程组 $A\boldsymbol{x}=\boldsymbol{c}$ 的变换．第一步，我们交换第 1 行和第 2 行，使得交换以后 $A(1,\ 1)\neq 0$．第二步，我们从第 2 行中减去第 1 行的 $A(2, 1)/A(1, 1)$倍，从第 3 行中减去第 1 行的 $A(3, 1)/A(1, 1)$倍．在这一步变换以后，$A(2, 1)=A(3, 1)=0$．第三步，保证 $A(2, 2)\neq 0$，否则需要交换第 2 行与第 3 行．然后从第 3 行中减去 $A(3, 2)/A(2, 2)$倍的第 2 行，使得 $A(3, 2)=0$．做完这一步以后，我们可以自下而上地求解方程组．从第三的式子可以解出 $x_3=12/(-4)=-3$．我们把这个值带入第二个等式 $2x_2+4x_3=14$，即 $2x_2+4(-3)=14$，所以 $x_2=13$．最后，第一个等式给出 $3x_1+6x_2+9x_3=12$，即 $3x_1+6(13)+9(-3)=12$，所以 $x_1=-13$．

为什么这种算法是可行的呢？关键在于，每一步中都有办法通过重新排列下方的

① B 的第 j 列定义为 $g(BE_j)$，其中向量 BE_j 的元素为 $BE_j(i)=1\{i=j\}$．

各行使得 $A(i, i) \neq 0$．如果在第 i 步有 $A(i, i) = 0$ 且 $A(j, i) = 0$ 对于所有 $j > i$ 成立，则这个算法就无法继续进行了．例如，我们有可能在其中的某一步将 $A\boldsymbol{x} = \boldsymbol{c}$ 化简成：

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 2 \\ 0 & 0 & -4 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} 4 \\ 7 \\ 12 \end{bmatrix}.$$

此时，我们会发现$-2\boldsymbol{v}_1+\boldsymbol{v}_2 = 0$，所以 A 的列向量之间不是线性无关的．但是通过前面的变换，列向量之间能否变为线性相关？我们知道采用行交换不会导致这个结果．下面我们说明从一行减去另一行的倍数也不会造成这样的情况．假设列向量 $\boldsymbol{v}_i$ 都是线性无关的．现在从第 j 行减去第 k 行的 α 倍，这相当于用 $\boldsymbol{v}_i(j) - \alpha \boldsymbol{v}_i(k)$ 替换 $\boldsymbol{v}_i(j)$ ．

如果有

$$\sum_i a_i \left[\boldsymbol{v}_i(j) - \alpha \boldsymbol{v}_i(k) \right] = 0 \text{ 且 } \sum_i a_i \boldsymbol{v}_i(r) = 0, \forall r \neq j,$$

则

$$\sum_i a_i \boldsymbol{v}_i(k) = 0.$$

那么用 α 乘以上式并与第一个等式相加，可以推出 $\sum_i a_i \boldsymbol{v}_i = 0$ ．也就是说，如果经过上述变换得到的矩阵 A 中行向量线性相关，那么最初矩阵对应的行向量就是彼此线性相关的．由此，我们证明了这种算法是切实可用的．

B.5 特征值

特征值、特征向量的概念对于矩阵的表示具有十分关键的作用，也说明了矩阵是如何对向量实现变换的．

设 $A \in \Re^{n \times n}$ ．如果

$$A\boldsymbol{v} = \lambda \boldsymbol{v},$$

则称一个非零向量 $\boldsymbol{v} \in \Re^n$ 为 A 的相应于特征值 λ 的右特征向量．

这时，

$$(A - \lambda \boldsymbol{I})\boldsymbol{v} = \boldsymbol{0} ,$$

这表明矩阵 $A - \lambda \boldsymbol{I}$ 没有逆矩阵．事实上，如果它有逆矩阵，可以推出

$$\boldsymbol{v} = (A - \lambda \boldsymbol{I})^{-1}(A - \lambda \boldsymbol{I})\boldsymbol{v} = (A - \lambda \boldsymbol{I})^{-1}\boldsymbol{0} = \boldsymbol{0} .$$

所以， $A - \lambda \boldsymbol{I}$ 的行列式一定等于 0.

因此，矩阵 A 的特征值就是多项式 $\det(A - \lambda \boldsymbol{I})$ 零点的值．所以，我们可以通过这些零点来找出特征值．对于每一个特征值 λ ，都可以通过解线性方程组 $A\boldsymbol{v} = \lambda \boldsymbol{v}$ 来找

到相应的特征向量．注意，任何一个特征向量的倍数同样是特征向量：如果 $\boldsymbol{v}$ 是相应于特征值 λ 的特征向量，那么 $\alpha\boldsymbol{v}(\alpha \neq 0)$ 也是相应于特征值 λ 的特征向量．

B.6 表示方式

矩阵有一些特殊的表示形式，能够便于求出矩阵的幂以及矩阵作用在向量上的效果．

设 $A \in \Re^{n \times n}$，我们用对角阵 $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$ 解释其 2 种表示形式，其对角线上元素等于 λ_1，非对角线上元素为 0，即 $\Lambda(i, j) = \lambda_i 1\{i = j\}$．

我们将阐述矩阵的 3 种表示形式．第一种，若 A 的特征值 λ_1 互不相同，则有

$$A = V\Lambda V^{-1},$$

这里矩阵 V 的列向量是 A 的特征向量．这被称为特征值分解．这种分解方法的好处在于它提供了一个简便的方法来计算矩阵 A 的幂．具体来说，我们有

$$A^k = V\Lambda^k V^{-1},$$

和 $\Lambda^k = \mathrm{diag}\left(\lambda_1^k, \cdots, \lambda_n^k\right)$．

第二，若 A 的特征值不是互不相同的，则 A 可以表示为

$$A = VJV^{-1},$$

其中 J 为 Jordan 矩阵．这里的 J^k 同样是易于计算的．

第三，在第二种情况里，我们也可以将 A 写成

$$A = U\Lambda V',$$

其中 $U'U = V'V = \boldsymbol{I}$．这被称为奇异值分解．在本书 6.4 节介绍联合高斯随机变量的表示以及 10.3 节介绍推荐机制时，我们会解释这种分解方式的应用．

B.6.1 互不相同的特征值

设 A 有 n 个特征向量 $\boldsymbol{v}_i$，对应的特征值 λ_1 互不相同，下面证明这些向量是线性无关的．

假设 $\boldsymbol{v}_i$ 并不是线性无关的．令 S 表示$\{1, \cdots, n\}$的子集中满足

$$\sum_{i \in S} b_i \boldsymbol{v}_i = \boldsymbol{0},$$

且含有向量个数最少的子集．对所有的 $i \in S$，有 $b_i \neq 0$．考虑 S 中的元素 j，我们得到：

$$\begin{aligned}\mathbf{0} &= A\left(\sum_{i\in S} b_i \boldsymbol{v}_i\right) = \sum_{i\in S} b_i \lambda_i \boldsymbol{v}_i \\ &= \sum_{i\in S} b_i \lambda_i \boldsymbol{v}_i - \lambda_j \left(\sum_{i\in S} b_i \boldsymbol{v}_i\right) \\ &= \sum_{i\in S} b_i (\lambda_i - \lambda_j) \boldsymbol{v}_i \\ &= \sum_{i\in S\setminus\{j\}} c_i \boldsymbol{v}_i .\end{aligned}$$

其中 $c_j = b_i(\lambda_i - \lambda_j)$. 这与子集 S 中使得 $\boldsymbol{v}_i$ 的线性组合为 0 的元素个数最少相矛盾. 因此，这些向量是线性无关的.

我们接下来定义矩阵 $V = [\boldsymbol{v}_1|\cdots|\boldsymbol{v}_n]$，它的列向量是 $\boldsymbol{v}_i$. 则

$$AV = [\lambda_1 \boldsymbol{v}_1 | \cdots | \lambda_n \boldsymbol{v}_n] = V\Lambda,$$

其中 $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_n)$.

对于矩阵 V，只有当 $\boldsymbol{a} = 0$ 时才有 $V\boldsymbol{a} = 0$，所以 V 存在逆矩阵 V^{-1}. 因此

$$A = V\Lambda V^{-1}.$$

那么，一个矩阵 A 什么时候才有互不相同的特征值呢？我们已经说过特征值就是下式的根：

$$\det(A - \lambda \boldsymbol{I}) = 0 .$$

等式左侧是 n 阶的多项式. 根据代数基本定理，这个多项式可以分解为

$$a(\lambda - \lambda_1) \times \cdots \times (\lambda - \lambda_n).$$

所以，如果某个 λ_1 是重根，就意味着在多项式 $\det(A - \lambda \boldsymbol{I})$ 中 λ 不同幂次的系数满足一定的代数关系. 举例来说，如果 $b^2 = 4ac$，则 $a\lambda^2 + b\lambda + c = 0$ 有重根. 如果矩阵 $A \in \Re^{n\times n}$ 是通过“随机”选取矩阵元素而得到的，例如从[−10, 10]中选取，则矩阵 A 的特征值互不相同的概率是 1.

虽然如此，我们依然需要研究特征值相同时该如何处理.

B.6.2 Jordan标准型

考虑以下 Jordan 标准型的例子：

$$J(4,\lambda) = \begin{bmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \end{bmatrix}.$$

通过归纳，我们可以得到

$$J(4,\lambda)^k=\begin{bmatrix}\lambda^k & \binom{k}{1}\lambda^{k-1} & \binom{k}{2}\lambda^{k-2} & \binom{k}{3}\lambda^{k-3}\\ 0 & \lambda^k & \binom{k}{1}\lambda^{k-1} & \binom{k}{2}\lambda^{k-2}\\ 0 & 0 & \lambda^k & \binom{k}{1}\lambda^{k-1}\end{bmatrix}.$$

类似地，对于$m\geqslant 1$我们可以定义$J(m,\lambda)$，读者可以猜出$J(m,\lambda)^k$的形式．下一步，考虑块对角阵

$$J=\begin{bmatrix}J(4,a) & 0\\ 0 & J(3,b)\end{bmatrix}=\begin{bmatrix}a & 1 & 0 & 0 & 0 & 0 & 0\\ 0 & a & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & a & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & a & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & b & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & b & 1\end{bmatrix}.$$

我们可以验证

$$J^k=\begin{bmatrix}J(4,a)^k & 0\\ 0 & J(3,b)^k\end{bmatrix}.$$

因此，类似形式矩阵的幂次很容易计算．我们把这样的结构叫作 Jordan 型．

设矩阵A可以写作

$$A=VJV^{-1}, \tag{B.6}$$

则$A^k=VJ^kV^{-1}$也是容易计算得到的．

事实上，任意的方阵A都可以写成上述形式，其中J具有块 Jordan 型的形式．它的 Jordan 块为$J(n_i,\lambda_i)$，其中每一个λ_1都是A的特征值，n_i是相应特征值的重数，即

$$\det(A-\lambda\boldsymbol{I})=a(\lambda-\lambda_1)^{n_1}\cdots(\lambda-\lambda_m)^{n_m},$$

其中a为常数．分解方法（B.6）称为A的 Jordan 标准型．

B.6.3 奇异值分解

现在讨论方阵$A\in\Re^{n\times n}$的奇异值分解．设$m\leqslant n$为A中线性无关列数的最大值（除非A是满秩的，否则$m<n$），那么

$$A=U\Lambda V'.$$

其中$U'U=V'V=\boldsymbol{I}$，$\Lambda=\mathrm{diag}(\sigma_1,\cdots,\sigma_m,0,\cdots,0)$．这里的$\sigma_i$表示$AA'$的正特征值的平方根．$V$的列向量是$A'A$的特征向量．$U$的列向量是$AA'$的特征向量．

举例来说，设

$$A = \begin{bmatrix} 1 & \sqrt{3} \\ 2 & 0 \end{bmatrix},$$

则

$$AA' = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}, \quad A'A = \begin{bmatrix} 5 & \sqrt{3} \\ \sqrt{3} & 5 \end{bmatrix}.$$

AA'的特征值为$\sigma_1^2 = 6$和$\sigma_2^2 = 2$．根据定义，我们可以得到 U 和 V:

$$U = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & -1 \\ -1 & 1 \end{bmatrix}, \quad V = \frac{1}{2} \begin{bmatrix} -\sqrt{3} & 1 \\ -1 & -\sqrt{3} \end{bmatrix}.$$

这里对特征向量进行了归一化，使得 $U'U = V'V = \boldsymbol{I}$. 于是，$A$ 的奇异值分解为

$$A = U\Lambda V' = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{6} & 0 \\ 0 & \sqrt{2} \end{bmatrix} \frac{1}{2} \begin{bmatrix} -\sqrt{3} & 1 \\ -1 & -\sqrt{3} \end{bmatrix}.$$

设 $U = [\boldsymbol{u}_1|\boldsymbol{u}_2]$以及 $V = [\boldsymbol{v}_1|\boldsymbol{v}_2]$. 注意到 $AV = U\Lambda$，因此 $A\boldsymbol{v}_i = \sigma_i \boldsymbol{u}_i$．这样，互相正交的向量 $\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$ 就与正交的向量 $\sigma_1 \boldsymbol{u}_1$ 和 $\sigma_2 \boldsymbol{u}_2$ 对应起来了.

考虑如下定义的单位圆：

$$\mathcal{C} = \{(x_1, x_2) \in \Re^2 \mid x_1^2 + x_2^2 = 1\}.$$

注意到一个等效的表达方式是

$$\mathcal{C} = \{\alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2 \mid \alpha_1^2 + \alpha_2^2 = 1\}.$$

原因是，由于 $V'V=\boldsymbol{I}$，我们有

$$(\alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2)'(\alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2) = \alpha_1^2 + \alpha_2^2.$$

通过矩阵 A，圆 $\mathcal{C}$ 可以与新的集合 $A\mathcal{C}$ 对应起来，其中 $A\mathcal{C}$ 定义为

$$A\mathcal{C} = \{A\boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{C}\}.$$

因此，

$$\begin{aligned} A\mathcal{C} &= \{A(\alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2) \mid \alpha_1^2 + \alpha_2^2 = 1\} \\ &= \{\alpha_1 \sigma_1 \boldsymbol{u}_1 + \alpha_2 \sigma_2 \boldsymbol{u}_2) \mid \alpha_1^2 + \alpha_2^2 = 1\}. \end{aligned}$$

这个集合是一个椭圆，它的轴是 $\boldsymbol{u}_1$ 和 $\boldsymbol{u}_2$，如图 B-1 所示.

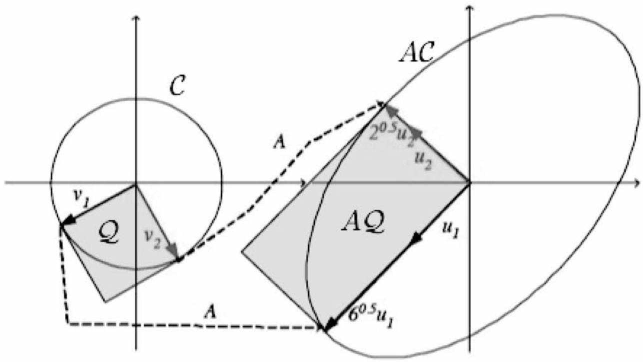


图 B-1　A 的作用是旋转和拉伸

再举一例．考虑 A 作用在方阵 $\mathcal{Q}$ 上的情况，

$$\mathcal{Q}=\{\alpha_1\boldsymbol{v}_1+\alpha_2\boldsymbol{v}_2|\ \alpha_1,\alpha_2\in[0,1]\}.$$

如图 B-1 所示，A 将这个集合变为 $A\mathcal{Q}$，

$$A\mathcal{Q}=\{\alpha_1\sigma_1\boldsymbol{u}_1+\alpha_2\sigma_2\boldsymbol{u}_2|\ \alpha_1,\alpha_2\in[0,1]\}.$$

$A\mathcal{Q}$ 的面积为 $\alpha_1\alpha_2=|\det(A)|$，也就是 $\mathcal{Q}$ 的面积乘以 $|\det(A)|$．

B.7　参考资料

线性代数有很多优秀的参考书籍，同时在网络上找到也可以找到大量资料．Strang 的 *Introduction to Linear Algebra* 一书很好地描述了线性代数的有关性质，内容深度适合本书的读者．

B.8　练习

1. 证明若 $ad\neq bc$，则

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1}=\frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

2. 使用 Matlab 计算下面矩阵 A 的奇异值分解，画出 A 作用在单位圆上时 AC 表示的图形．　C.1

$$A=\begin{bmatrix} 2 & 3 \\ 1 & 0 \end{bmatrix}.$$

3. 使用 Matlab 找到下面矩阵 B 的特征值分解，写出 B^k．　C.1

$$B = \begin{bmatrix} 0.05 & -0.35 \\ -0.35 & 0.05 \end{bmatrix}.$$

4. 对上一题的 B，设 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 是已知向量，用特征值分解的性质给出当 k 足够大时 $\boldsymbol{a}'B^k\boldsymbol{b}$ 的近似值.

C.1

5. 考虑下面的矩阵 C：

$$C = \begin{bmatrix} 1 & -3 & -2 \\ -1 & 1 & -1 \\ 2 & 4 & 5 \end{bmatrix}.$$

使用 Matlab 找到 C 的 Jordan 标准型.

提示：查看 Matlab 的文件.

6. 利用 C 的 Jordan 标准型，直接写出 C^k .

附录C

Matlab

我们使用 Matlab 软件作为仿真和计算的工具．这篇附录整理了一些标准的代码．

作者自认为不是书写编程知识的最佳人选，所以请读者不要把这里给出的例子作为编程的模板．给出这些例子是为了让读者节省一些时间．为了增强代码的可读性，书中的很多代码可能不是最简化的，希望读者理解．

C.1　矩阵

B.2

Matlab 针对矩阵进行运算．

```
A=[1,2,3;4,5,6]
```

会产生这样的矩阵：

$$A=\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\end{bmatrix}.$$

同时，代码

```
A*B,A',A^(-1),det(A),eig(A),[V,D]=eig(A)
```

可以依次产生矩阵 A 和 B 的乘积矩阵，A 的转置，A 的逆矩阵，A 的行列式，由 A 的特征值组成的列向量，以 A 的特征向量为列向量的矩阵 V，以及以 A 的特征值为对角线上元素的对角阵 D．也就是说，$AV=VD$ 且 $A=VDV^{-1}$．例如，

```
A=[1,2;3,4];
[V,D]=eig(A)
```

会产生

```
V =

   -0.8246   -0.4160
    0.5658   -0.9094

D =
```

```
   -0.3723         0
         0    5.3723
```

再举一例，对于随机矩阵

B.5 和 B.6

```
P=[0,0.3,0.7;0,0.4,0.6;1,0,0];
```

我们有

```
>> eig(P)

ans =

   -0.7359
    1.0000
    0.1359
```

这验证了一个不可约、非周期的马尔可夫链对应的随机矩阵有且只有一个等于 1 的特征值，且其他特征值的绝对值严格小于 1（Perron-Frobenius 定理）.

同时，我们有

```
>> [V,D]=eig(P)

V =

   -0.5454    0.5774    0.0547
   -0.3915    0.5774   -0.9139
    0.7411    0.5774    0.4023


D =

   -0.7359         0         0
         0    1.0000         0
         0         0    0.1359
```

可以验证

$$P = VDV^{-1}.$$

事实上，

```
>> V*D*V^(-1)

ans =

    0.0000    0.3000    0.7000
    0.0000    0.4000    0.6000
    1.0000    0.0000    0.0000
```

因此，$V^{-1}P = DV^{-1}$. 这意味着如果 $\boldsymbol{w}_k$ 是 V^{-1} 的第 k 行，则 $\boldsymbol{w}_k = \lambda_k \boldsymbol{w}_k$，其中 λ_k 是第 k 个特征值. 所以，

$$\boldsymbol{w}_k P^n = \lambda_k^n \boldsymbol{w}_k.$$

如果我们把一个向量 π 写成

$$\pi = a_1\boldsymbol{w}_1 + a_2\boldsymbol{w}_2 + a_3\boldsymbol{w}_3 \text{ ,}$$

则

$$\pi P^n = \sum_{k=1}^{3} a_k \lambda_k^n \boldsymbol{w}_k \to a_2\boldsymbol{w}_2.$$

因为 $\lambda_2 = 1$ 且 $|\lambda_1| < 1$，$|\lambda_3| < 1$．注意，这里可以采用这种表达形式是因为特征值互不相同，否则矩阵 D 需要用块 Jordan 矩阵来代替（参阅 B.6 节）.

如果随机矩阵不可约，且不是非周期的，情况会有所不同．例如，

```
>> P=[0,1;1,0];
>> [V,D]=eig(P)

V =

   -0.7071    0.7071
    0.7071    0.7071


D =

    -1     0
     0     1
```

还有，

```
A(m;:), A (:, n)
```

分别表示 A 的第 m 行和第 n 列．例如，

$A(1,:) = [1,2,3]$ 和 $A(:,2) = [2;5]$.

C.2　随机变量

A.2和A.3

Matlab 有一个函数 `rand(1)`，可以返回一个均匀分布在[0, 1]上的随机变量．连续调用这个函数，可以产生满足均匀分布的独立同分布随机变量.

因此，如果我们希望得到一个服从两点分布 $B(p)$的随机变量 X，可以用以下代码：

```
if rand(1)<p
    X=1;
else
    X=0;
end
```

采用 Matlab 可以很容易地产生更复杂的分布：

```
random('bino',n,p)
```

返回的是服从 $B(n, p)$分布的随机变量．如果 $n=1$，则

```
random('bino',1,p)
```

返回的是服从 $B(p)$的随机变量．同时，

```
random('exp',λ),random('geo',p),
random('Poisson',λ),random('Normal',μ,σ)
```

各自返回我们希望得到的随机变量．我们也可以通过加上参数 m 和 n 来产生由这些变量组成的 $m\times n$ 阵列．如

```
random('Normal',0,1,3,4)
```

返回了 3×4 的阵列，其中的每个元素都是相互独立的服从 $\mathcal{N}(0, 1)$分布的随机变量．

要产生一个取值为 $k(k=1, \cdots, K)$的概率是 $P(k)$的随机变量，可以用下述方法：

```
function T=discrete(P)
% 该函数产生一个概率服从Pr(x=k)=P(k)/sum(P)分布的随机变量
% P=[P(1),P(2),…,P(k)],其中P(k)均为非负
Pnorm=[ 0 P]/sum(P); %见下方说明(1)
Pcum=cumsum(Pnorm); %(2)
R=rand(1);
[~,T] = histc(R,Pcum); %(3)
```

现在对上述代码进行说明．(1)为向量 P 加一个 0，并且使其归一化．(2)建立 P 的累积分布函数使得第一个元素依然是 0，而最后一个元素是 1．也就是说，$Pcum(k)=\sum_{i=1}^{k} Pnorm(i)$．(3)$R$ 均匀分布在 0 到 1 之间，可以认为直方图函数的功能是求在 R 处测量的累积分布函数的拟逆．事实上，$T=\sup_i Pcum(i)\leqslant R$．这解释了为什么要让 P 的开头为 0.

例如，

```
P=[1,2,3,4];
discrete(P)
```

产生了一个值，取值为 1 的概率是 0.1，取值为 2 的概率是 0.2，以此类推.

C.3 马尔可夫链

1.2

设 P 是一个 $N\times N$ 的随机矩阵.

C.3.1 仿真

我们希望得到一个转移概率矩阵为 P 的马尔可夫链 $\boldsymbol{X}=\{X(m), m=1,\cdots,M\}$．已知 $X(n)$时，可以通过 $discrete(P(X(n), :))$来产生 $X(n+1)$．那么，$X(1)=A$，通过 M 步来仿真 $\boldsymbol{X}$ 的代码：

```
function simMC(M,A,P)
%对一个初始状态为A、转移矩阵为P的马尔可夫链进行M步模拟
X = zeros(1,M);
```

```
X(1) = A;
for m = 1:M-1
    X(m+1) = discrete(P(X(m)),:));
end
plot(X,'.-b')
```

图 C-1 给出了用如下代码得到的仿真结果：

```
M=100; A=1; P=[1-a,a;a,1-a];
simMC(M,A,P)
```

其中为 a = 0.1、0.2 和 0.4．图中的 Y 轴都被重新标注了，因为 Matlab 代码产生了一个在{1, 2, …, N}上的马尔可夫链，而不是在{0,1}上的．

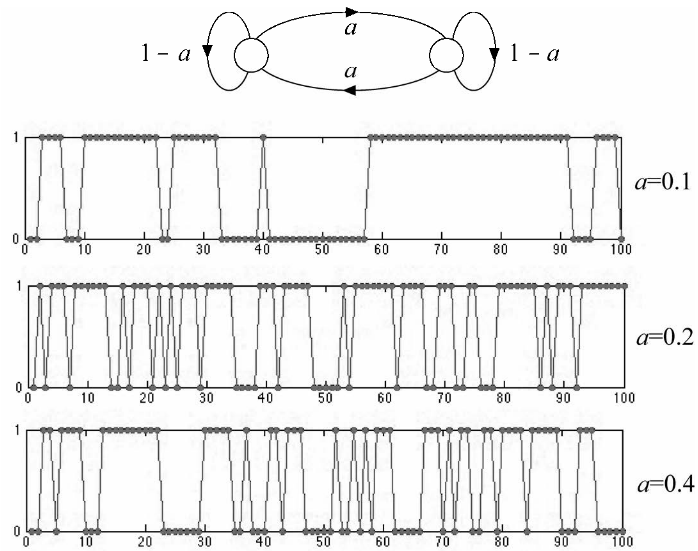


图 C-1　对于一个 2 态马尔可夫链的仿真

再举一例，图 C-2 给出了对于 5 态马尔可夫链的仿真（这对应于前面网页搜索的例子）．用到的代码如下：

```
M=100; A=1;
P=[0,0.5,0,0.5,0;0,0,1,0,0;1,0,0,0,0;1/3,1/3,0,0,1/3;0,0.5,0.5,0,0];
simMC(M,A,P)
```

同样地，我们重新标注 Y 轴，把{1, 2, 3, 4, 5}改成{A, B, C, D, E}．

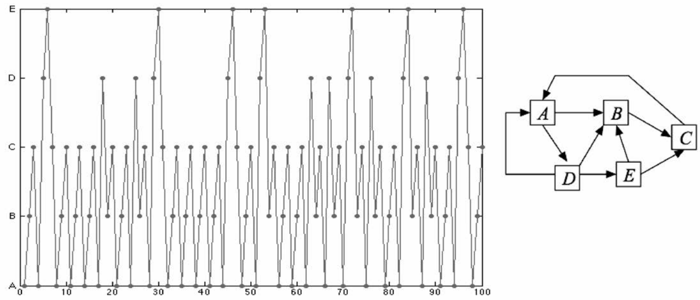


图 C-2　对于一个 5 态马尔可夫链的仿真

C.3.2　稳态分布

设 P 是不可约的．那马尔可夫链的平衡方程为

$$\pi = \pi P .$$

由定理 1.2，我们知道对于这些方程有唯一的解 π，满足所有元素之和为 1．可以将方程组重写为

$$\pi(P-I) = \mathbf{0} ,$$

其中 I 为单位阵．设 $\boldsymbol{c}_1,\cdots,\boldsymbol{c}_N$ 是矩阵 $P–I$ 的列．那么方程组就是

$$\pi\boldsymbol{c}_k = 0, k = 1,\cdots, N.$$

这些方程不是相互独立的．事实上，$P–I$ 的各行之和为 $\mathbf{0}$ 向量，因为

$$(P-I)\mathbf{1} = P\mathbf{1} - \mathbf{1} = \mathbf{0} ,$$

其中 $\mathbf{1}$ 为每个元素均为 1 的 N 维列向量．因此，如果对于 $k = 1,\cdots, N-1$ 都有 $\pi\boldsymbol{c}_k = 0$，则

$$\pi\boldsymbol{c}_N = \pi(\mathbf{0} - \boldsymbol{c}_1 - \cdots - \boldsymbol{c}_{N-1}) = 0.$$

所以，最后一个等式是多余的．我们把它替换为

$$\pi\mathbf{1} = 1 .$$

也就是考虑以下方程组：

$$\pi[\boldsymbol{c}_1\boldsymbol{c}_2\cdots\boldsymbol{c}_{N-1}\mathbf{1}] = [0,0,\cdots,0,1].$$

因为有唯一解 π，所以这些方程一定是线性无关的．那么，我们有

$$\pi = [0,0,\cdots,0,1][\boldsymbol{c}_1\boldsymbol{c}_2\cdots\boldsymbol{c}_{N-1}\mathbf{1}]^{-1} .$$

举一个简单的例子，设

$$P=\begin{bmatrix}1-a & a\\ b & 1-b\end{bmatrix},$$

则

$$P-I=\begin{bmatrix}-a & a\\ b & -b\end{bmatrix}.$$

前面讲过

$$\begin{bmatrix}a & b\\ c & d\end{bmatrix}^{-1}=\frac{1}{ad-bc}\begin{bmatrix}d & -b\\ -c & a\end{bmatrix},$$

所以，关于 π 的方程就是

$$\pi=[0,1]\begin{bmatrix}-a & 1\\ b & 1\end{bmatrix}^{-1}=[0,1]\frac{1}{a+b}\begin{bmatrix}-1 & 1\\ b & a\end{bmatrix}=\left[\frac{b}{a+b},\frac{a}{a+b}\right].$$

下列函数计算了 P 的稳态分布：

```
function [INVD]=invdist(P)
%该函数计算给定随机转移矩阵P下的马尔可夫链的稳态分布
N = size(P,2);
B = (P - eye(N));
B(1:N,N) = ones(N,1);
o = zeros(1,N);
o(N) = 1;
INVD = o*inv(B);
```

举个例子，

```
P=[0,0.3,0.7;0,0.4,0.6;1,0,0];
invdist(P)
```

会产生一个不变分布：

```
ans =

    0.4000    0.2000    0.4000
```

C.4 置信区间

现在讨论置信区间．设 $\{X_n, n\geqslant 1\}$ 是均值为 μ ，方差为 σ^2 的独立同分布随机变量．在 3.4 节中我们解释过，当 n 足够大时，有 95%置信区间的均值 μ 为

3.4

$$\left[\mu_n-\frac{2\sigma}{\sqrt{n}},\ \mu_n+\frac{2\sigma}{\sqrt{n}}\right],$$

其中

$$\mu_n = \frac{X_1 + \cdots + X_n}{n}$$

为样本的均值．我们通过抛掷硬币来检验这一置信区间．

对于抛掷硬币，也就是服从两点分布 $B(p)$的随机变量，我们有

$$\mathrm{var}(X_n) = p(1-p) \leqslant 1/4 .$$

然后，将 1/2 作为标准偏差σ 的上界．σ 是针对如下给定的置信区间而言的：

$$\left[\mu_n - \frac{1}{\sqrt{n}},\ \mu_n + \frac{1}{\sqrt{n}}\right].$$

下面的代码展示了独立同分布的 $B(p)$随机变量 $X(1)$, …, $X(N)$并计算了样本均值．代码还计算了置信区间的界限．图 C-3 展示了计算结果．

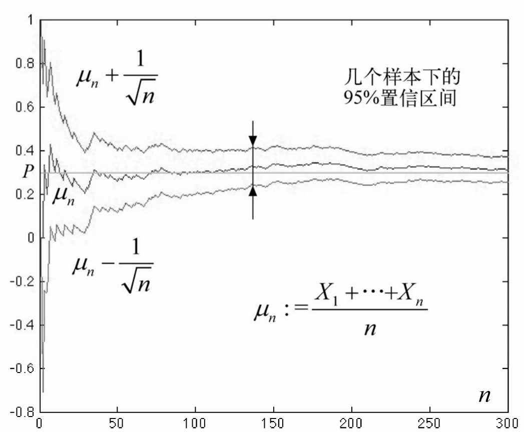


图 C-3　通过方差上界得到关于硬币偏移值的置信区间（另见彩插）

```
%抛掷硬币N次，得到N个概率为P的伯努利随机变量
%根据采样平均以及方差不大于1/4的条件构建置信区间
N = 300;
p = 0.3;
Y = zeros(1,N);
U = ones(1,N);
M = p*U;
L = zeros(1,N);
Y(1) = (rand(1)<p);
for j = 2:N
Y(j) = ((j-1)*Y(j-1)+(rand(1)<p))/j;
b = 1/sqrt(j);
U(j) = Y(j) + b;
L(j) = Y(j) - b;
end
plot([Y;L;U;M]')
```

除了采用置信区间的上界来得到置信区间的范围，也可以采用样本的标准偏差.这时，置信区间为

$$\left[\mu_n-\frac{2\sigma_n}{\sqrt{n}},\mu_n+\frac{2\sigma_n}{\sqrt{n}}\right],$$

其中

$$\sigma_n^2=\frac{n}{n-1}\left\{\frac{1}{n}\sum_{m=1}^{n}X(m)^2-\mu_n^2\right\}$$

是无偏的样本方差（参阅 3.4 节）. 下面是修改后的代码．图 C-4 展示了结果.

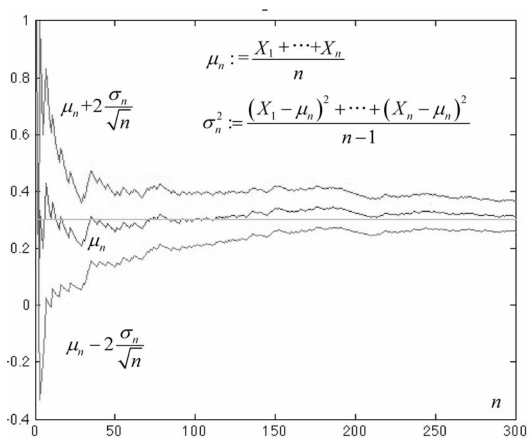


图 C-4 通过采用样本方差得到硬币偏移的置信区间（另见彩插）

起初，采用样本标准偏差 σ_n 的效果并不是很好．不过，随着 n 的增大，采用这个估计值得到的结果比采用上界的值更准确，能产生更小的置信区间.

```
% 产生N个概率为P的伯努利随机变量
% 根据采样平均与采样方差构建置信区间
N = 300;
Y = zeros(1,N); %样本平均
U = ones(1,N); %置信区间上界
M = 7*U;        %真实期望
L = zeros(1,N); %置信区间下界
S = U; %样本二阶矩
Y(1) = (rand(1)<p); %B(p)
S(1) = Y(1)^2;
for j = 2:N
    R = (rand(1)<p); %B(p)
    Y(j) = ((j-1)*Y(j-1)+R)/j;  %Y(j)=前j个随机变量的样本平均
    S(j) = ((j-1)*S(j-1)+R^2)/j; %S(j)=样本二阶矩
    b = 2/sqrt(j); %调整sigma以获得95%的置信区间
    sigma = sqrt((j/(j-1))*(S(j)-Y(j)^2)); %样本标准方差
             % 这里采用 sqrt(j/(j-1)) 保证无偏估计
```

```
    U(j) = Y(j) + b*sigma;
    L(j) = Y(j) - b*sigma;
end
plot([Y;L;U;M]')
```

上面的代码也可以用于其他的分布．这里给出一个指数分布的例子．图 C-5 给出了运行结果．

```
% 产生N个期望值为7的随机变量
% 根据样本平均与样本方差构建置信区间
N = 300;
Y = zeros(1,N); %样本平均
U = ones(1,N); %置信区间上界
M = 7*U;      %真实期望
L = zeros(1,N); %置信区间下界
S = U; %样本二阶矩
Y(1) = random('exp',7); %期望为7的指数分布
S(1) = V(1)^2;
for j = 2:N
    R = random('exp',7); %期望为7的指数分布
    Y(j) = ((j-1)*Y(j-1)+R)/j;  %前j个随机变量的样本随机
    S(j) = ((j-1)*S(j-1)+R^2)/j; %样本二阶矩
    B = 2/sqrt(j);  %调整sigma以获得95%的置信区间
    sigma = sqrt((j/(j-1))*(S(j)-Y(j)^2)); %样本标准方差
              % 这里采用 sqrt(j/(j-1)) 保证无偏估计
    U(j) = Y(j) + b*sigma;
    L(j) = Y(j) - b*sigma;
end
plot([Y;L;U;M]')
```

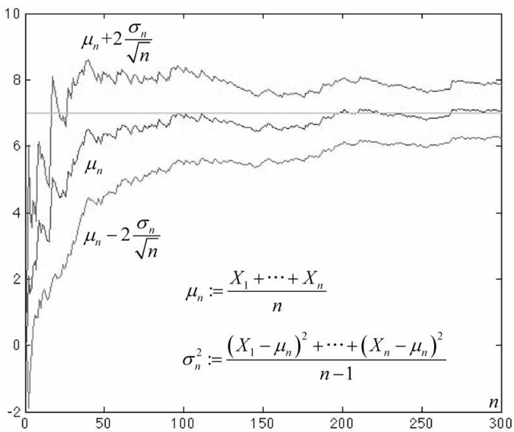


图 C-5　采用样本方差得到指数分布均值的置信区间（另见彩插）

C.5　数字链路

5.3

在 5.3 节中，我们研究了信道中含有加性高斯噪声时对比特信息接收的影响．这

一节中，我们用 Simulink 来探索这些系统.

要用 Simulink 工具，只需在 Matlab 的命令窗口里输入 simulink，然后选择“file →new model”，就可以使用模块库来组成通信系统. Simulink 有点像电子游戏，可以用来尝试不同的系统模型和仿真实验.

图 C-6 给出了一个服从独立同分布 $B(0.2)$ 的二进制信源. 信号经过一个满足 $N(0, 0.1)$ 的独立同分布高斯噪声信道，然后通过 $MAP[X|Y] = 1\{Y \geqslant 0.37\}$ 解码. 图中给出了信源波形、接收到的信号、最大后验概率以及误差.

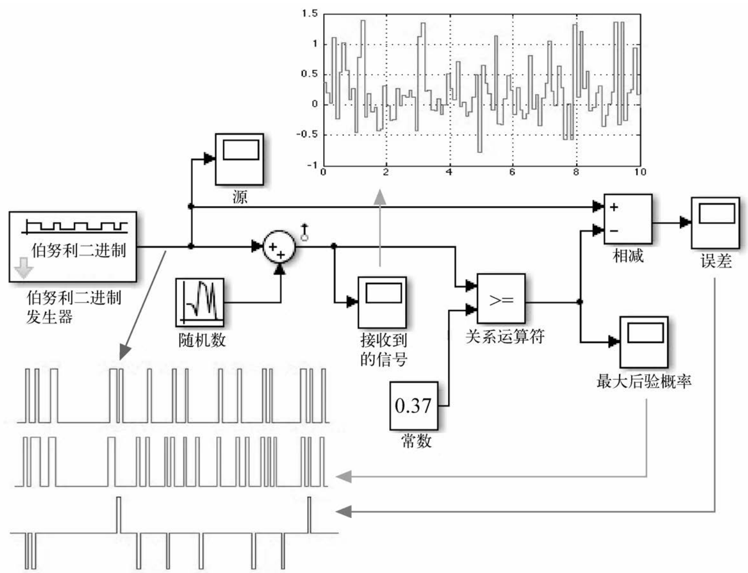


图 C-6　用 Simulink 探索加性高斯白噪声信道

C.6　LDPC 码

6.2

在 6.2 节中，我们介绍了 LDPC 码. 下面为其中置信传播解码的例子给出一个比较简略的 Matlab 实现方法.

```
%LDPC 解码算法

H=[1,0,1,1,1,0,0,0;
   0,1,0,1,1,0,1,0;
   1,1,0,0,0,1,0,1;
   0,0,1,0,1,1,1,1]; % y=Hx中的矩阵
xtrue=[0,1,0,0,1,0,1,0]'; % x的真实值
y=mod(H*true,2); % y的真实值
x=mod(xtrue+[0,0,0,0,1,0,0,0]',2); %x+误差
epsilon=0.05; %二元对称信道误差概率
```

```
Steps=10; %算法的执行步数

[N,M]=size(H);
Num=H;  % 式(6.8)的分子
Den=H;  % 式(6.8)的分母
epsilonbar=1-epsilon;
xh=zeros(M,N,Steps);
%对于y(j),估计 x(xh(i,j)=猜测 P(x(i)=1)
yh=zeros(N,M,Steps); %用y估计x
%由y(j),(yh(j,i)=猜测 P(x(i)=1)

A=zeros(1,M);
D=A;

Xh=zeros(M,Steps); %在第S步时对x的估计
Xd=Xh; %用以绘制 Xh
Xf=Xh; %最终预测

for i=1:M
    for j=1:N
        xh(i,j,1)=epsilon+(1-2*epsilon)*x(i);
    end
end
for s=1:Steps-1
    for j=1:N %对每次 y(j) 更新
        for i=1:M % 对每次由式(6.7)得到的yh(j,i)
            yh(j,i,s)=(1-2*epsilon)*(1-2*y(j));

            for k=1:M
                if(H(j,k)==1)&&ne(k,i)
                    yh(j,i,s)=yh(j,i,s)*(1-2*xh(k,j,s));
                end
            end
            yh(j,i,s)=0.5-0.5*yh(j,i,s);
        end
    end
    for i=1:M % 对每次 x(i) 更新
        for j=1:N % 对每次由式(6.8)得到的xh(j,i)

            if(H(j,i)==1)
                Num(i,j)=epsilon+(1-2*epsilon)*x(i);
                Den(i,j)=1-Num(i,j);
                for k=1:N
                    if(H(k,i)==1)&&ne(k,j)
                        Num(i,j)=Num(i,j)*yh(k,i,s);
                        Den(i,j)=Den(i,j)*(1-yh(k,i,s));
                    end
                    xh(i,j,s+1)=Num(i,j)/(Num(i,j)+Den(i,j));
                end
            end
        end
        for i=1:M      %通过对式(6.9)取整估计x(i)
            A(i)=epsilon+(1-2*epsilon)*x(i);
            D(i)=1-A(i);
```

```
        for j=1:N
            if(H(j,i)==1)
                A(i)=A(i)*yh(j,i,s);
                D(i)=D(i)*(1-yh(j,i,s));
            end
        end
        Xh(i,s+1)=round(A(i)/(A(i)+D(i)));
        Xd(i,s+1)=Xh(i,s+1)+0.01*i;
    end
end
plot(Xd');
Xf=round(Xh(:,Steps))-xtrue
```

C.7　卡尔曼滤波器

7.6

卡尔曼滤波器的公式尽管看起来很可怕，但却可以很容易地用 Matlab 来实现. 下面是一个计算卡尔曼滤波器的例子，结果如图 C-7 所示.

```
% 卡尔曼滤波器：系统 X+ = AX + V, Y = CX + W ( X+ 代表 X(n+1))
%          \Sigma_V = Q^2, \Sigma_W = R^2
%    构造高斯噪声 V = normrnd(0, Q), W = normrnd(0, R)
%        其中Z为独立同分布正态分布变量 N(0, 1)
%  该滤波器为 is Xh+ = (A - KCA)Xh + KY (Xh 为估计值)
%                    K+ = SC'[CSC' + R^2)
%                     S = AHA' + Q^2      (H = \Sigma 为估计误差的协方差)
%                    H+ = (1 - KC)S

% 常数
A = [1, 1; 0, 1];
C = [1, 0];
Q = [1; 0];
R = 0.5;
SV = Q*Q';
SW = R*R';

N = 20; %步数

M = length(A);

%系统
X = zeros(M, N);
Xh = X;
H = zeros(M, M);
K = H;

X(:,1) = [0; 3]; %初始状态

for n = 1:N-1

% 系统方程
    V = normrnd(0,Q);
```

```
     W = normrnd(0,R);
     X(:,n+1) = A*X(:,n) + V;
     Y = C*X(:,n+1) + W;

%滤波器
    S = A*H*A' + SV;              %得到计算
    K = S*C'*(C*S*C'+SW)^(-1);
    H = (1 - K*C)*S;
    %估计更新
    Xh(:,n+1) = (A - K*C*A)*Xh(:,n) + K*Y;

end

%绘图

P = [X(2,:); Xh(2,:)]';

plot(P)
```

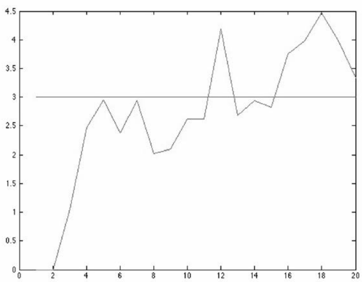


图 C-7　卡尔曼滤波器的仿真

Simulink 提供了一个卡尔曼滤波器的模块．图 C-8 展示了将卡尔曼滤波器应用到下述系统时的结果：

$$x(n+1)=0.9x(n)+v(n)$$
$$y(n)=x(n)+w(n),$$

其中 $v(n)$ 和 $w(n)$ 相互独立，并且 $v(n)=\mathcal{N}(0,\,0.2)$ ，$w(n)=\mathcal{N}(0,\,0.1)$ ．注意，为方便对比状态和滤波器的输出，我们同时定义了另一个等于 $x(n)$的标度．

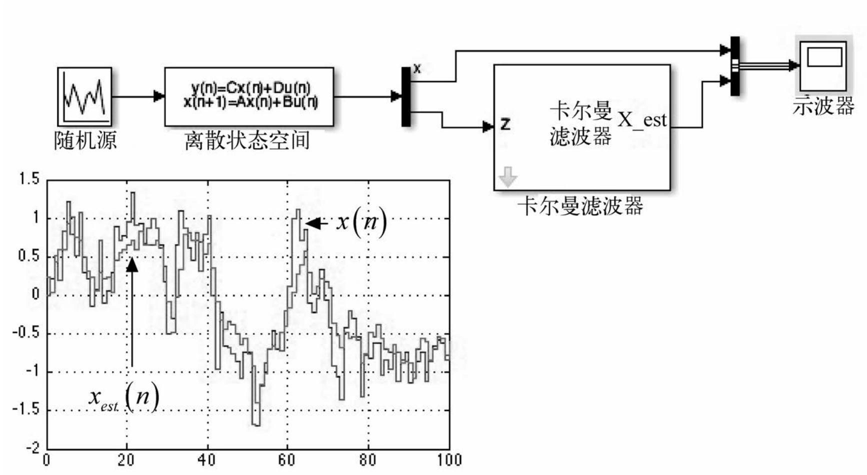


图 C-8 用 Simulink 模拟卡尔曼滤波器（另见彩插）

Simulink 模块的配置如下.

- 来源 Type=Gaussian；Mean=[0, 0]；Variance=[0.2, 0.1]
- 离散状态空间 A=0.9；B=[1,0]；C=[1;1]；D=[0,0;0,1]
- 卡尔曼滤波器 滤波器数量：1；初态：0；初始错误协方差：1

  状态转移矩阵：0.9；过程噪声协方差：0.1

  测量矩阵：1；测量噪声协方差：0.2

C.8 聚类

9.3

下面的程序优化了 9.3 节中具有 2 个高斯分布源的一维聚类算法.

C.8.1 硬期望最大化

这个程序会产生 100 个分数 $X(1), \cdots, X(100)$，它们独立同分布，并且以等概率服从 $N(a, s2)$或 $N(b, s2)$分布．然后程序执行 HEM 算法．这个算法在第 k 步时已有 $A(k)$和 $B(k)$作为 a 和 b 的估值．然后算法假设大于$(A(k)+B(k))/2$ 的 $X(n)$的值属于 A 等生，其余是 B 等生的分数．$A(k+1)$是 A 等生的平均分，$B(k+1)$是 B 等生的平均分．图 C-9 给出了结果.

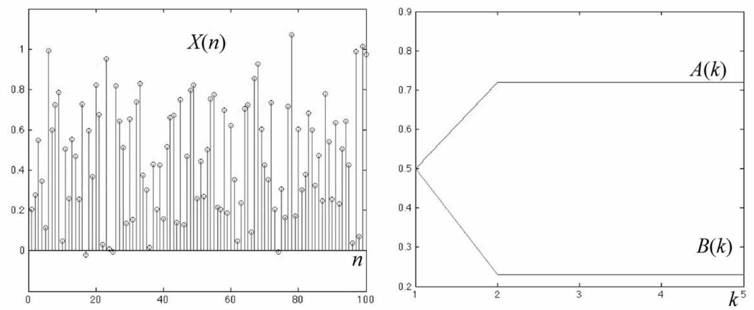


图 C-9　100 个数值（左图）以及均值的估计值的演变（右图）

```
% 采用硬期望最大化进行分类

a = 0.2;
b = 0.7; % 真实均值
s2 = 0.16; % 方差

N = 100; % 变量个数
X = zeros(1,N); % 变量

Steps = 5 ; % 步数
A = zeros(1, Steps);
B = A;
A(1) = 0.5;  % 初始估计
B(1) = 0.5;  % 对应均值为a和b的变量

for n = 1:N     % 生成100个分数值
    X(n) = a + (b - a)*random('bino',1, 0.5) + random('Normal',0, s2);
end

for k = 1:Steps-1
    AN = 0;
    M = (A(k) + B(k))/2;    % 中值
    for n = 1:N
        if X(n) > M         % 在这种情况下,判定n为A等生
            AN = AN + 1;    % 计算A等生数量
            A(k+1) = A(k+1) + X(n);   % A等生分数值总和
        else
            B(k+1) = B(k+1) + X(n);   % B等生分数值总和
        end
    end
    if AN > 0
        A(k+1) = A(k+1)/AN; % A等生样本平均
    end
    if AN < N
        B(k+1) = B(k+1)/(N - AN); % B等生样本平均
    end
end
G = [A; B];
plot(G')      % 绘制A和B
```

C.8.2 软期望最大化

软期望值最大化算法也是从猜测值(a_1, b_1)开始的．然后，算法会计算一个学生的成绩为 $X(n)$时，该学生是 A 等生的概率 $p(n)$．下一个猜测值 a_2 是用 $p(n)$对分数 $X(n)$进行加权平均得到的，b_2 的求解方法类似．即，

$$a_2 = \frac{\sum_n X(n)p(n)}{\sum_n p(n)}, \quad b_2 = \frac{\sum_n X(n)(1-(n))}{\sum_n (1-p(n))}.$$

图 C-10 展示了仿真结果．

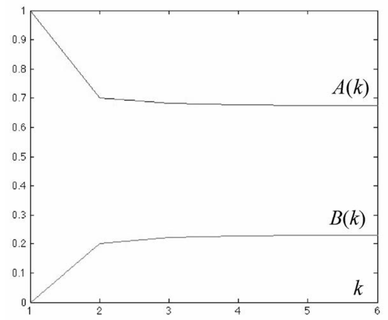


图 C-10　应用 SEM 均值的估计值的演变

```
% 采用软期望最大化进行分类

a = 0.7;
b = 0.2; % 真实均值
s2 = 0.04; % 方差

N = 100; % 变量个数
X = zeros(1,N); % 变量
Steps = 6 ; % 步数
A = ones(1, Steps);
B = zeros(1, Steps);
A(1) = 1; % 初始估计
B(1) = 0; % 对应均值为a和b的变量
for n = 1:N % 产生100个分数值
    X(n) = b + (a - b)*random('bino',1, 0.5) + random('Normal',0, s2);
end
for k = 1:Steps-1
    V = 0;
        W = 0;
        UA = 0;
        UB = 0;
    for n = 1:N
        p = 1 + exp(((X(n) - A(k))^2 - (X(n) - B(k))^2)/(2*s2));
        p = 1/p; % 为A等生出现概率
        V = V + X(n)*p;
```

```
            W = W + X(n)*(1 - p);
            UA = UA + p;
            UB = UB + 1 - p;
        end
        A(k+1) = V/UA;
        B(k+1) = W/UB;
    end
    G = [A; B];
    plot(G')     % 绘制A和B
```

当样本值的方差很大时，SEM 聚类算法无法收敛到真实的均值.

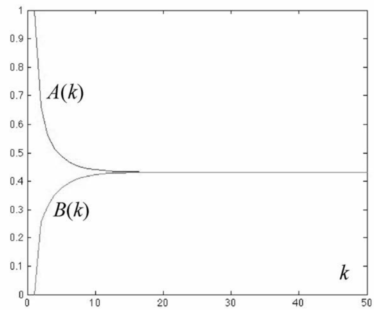


图 C-11　当样本值的方差很大时，SEM 聚类算法无法收敛到真实的均值

C.9　马尔可夫决策过程

11.4

下面是解决 11.4 节中“雇用帮工”问题的代码.

```
% 雇用一个帮工
K = 20; % 最大积压量
lambda = 0.5; % 到达率
mu = 0.6; % 服务率
alpha = 1.5; % 提速比
beta = 14; % 顾问价格
N = 200; % 时长
V = zeros(K+1, N+1);
a = V;
g = zeros(1,N+1);

for x = 1:K+1
    V(x, 1) = x ;
    a(x,1) = 0;
end

for m = 2:N+1
    for x = 1:K+1
        A = (1 - lambda)*(1 - mu)*V(x, m-1)
             + lambda*(1 - mu)*V(min(x+1, K), m-1)
             + (1 - lambda)*mu*V(max(x-1,1),m-1)
```

```
            + lambda*mu*V(x, m-1);
        B = beta + (1 - lambda)*(1 - mu*alpha)*V(x, m-1)
            + lambda*(1 - mu*alpha)*V(min(x+1, K), m-1)
            + (1 - lambda)*mu*alpha*V(max(x-1,1),m-1)
            + lambda*mu*alpha*V(x, m-1);
        % 无顾问
        V(x, m) = x + min(A, B);
        if A > B
            a(x, m) = 1;
        else
            a(x,m) = 0;
        end
    end
end

for m = 1:N+1
    z = find(a(:,N+ 2 - m));
    if isempty(z)
        g(m) = K+2;
    else
    g(m) = min(z);
    end
end

plot(g)
```

C.10 压缩感知

10.3

下面的 Matlab 代码可以解决 10.3 节中基于随机采样重建信号的问题.

```
%
Index = 10000; % 设置标度以绘制 s(t) 和重建的 r(t)
N = 100; % 样本空间大小
K = 15; % 样本数

s = zeros(1,Index); % 信号 s(t)
r = s; % 重建的信号 r(t)
Time = s; % 时间
phi = zeros(N,1);% 所有频率
x = zeros(1,N); % 估计系数
T = zeros(1, K); % 样本数量
b = T; % 样本
G = zeros(K,N); % 函数样本
beq = zeros(K,1); % 采样信号
B = [1, 2, 3]; % s(t)的系数
F = [1, 1.2, 1.6]'; % s(t)包含的频段

b = zeros(2*N, 1);

for k = 1:K
    T(k) = rand(1); % 采样时间
    beq(k) = B*sin(2*pi*F*T(k)) ; % 采样值
```

```
    for n = 1:N
        G(k, n) = sin(2*pi*(n/10)*T(k)); % g_n(t_k)
    end
end

f = [zeros(1, N), ones(1, N)]';
A = [eye(N), -eye(N); - eye(N), - eye(N)];
Aeq = [G,zeros(K,N)];

z = linprog(f,A,b,Aeq,beq); % 求解线性规划问题
for n = 1:N
    x(n) = z(n);
end


for n = 1:N
    phi(n) = n/10;
end

for i = 1:Index
    Time(i) = i/Index; % 时间从0到1
    s(i) = B*sin(2*pi*F*Time(i)); % 信号
    r(i) = x*sin(2*pi*phi*Time(i)); % 重建的信号
end

plot(Time,s,Time, r) % 绘制信号及其重建信号（蓝色，绿色）
```

C.11 参考资料

Matlab 附带的文件系统包含很多内容．网上也有很多可以参考的代码．

参考文献

N. Abramson. The ALOHA System - Another Alternative for Computer Communications. *Proc. 1970 Fall Joint Computer Conference*, 1970.

D. Aldous. *David Aldous's Home Page*. URL http://www.stat.berkeley.edu/~aldous/.

D. Bertsekas. Dynamic Programming and Optimal Control. Athena, 2005.

D. Bertsekas and J. Tsitsiklis. *Distributed and Parallel Computation: Numerical Methods*. Prentice-Hall, 1989.

D. Bertsekas and J. Tsitsiklis. *Introduction to Probability*. Athena, 2008.

P. Billingsley. *Probability and Measure, Third Edition*. Wiley, 2012.

P. Bremaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, 2008.

R.G. Brown and P.Y.C. Hwang. *Introduction to Random Signals and Applied Kalman Filtering*. Wiley, 1996.

E.J. Candes and B. Recht. Exact matrix completion via convex optimization. *Foundation of Computational Mathematics*, 9:717–772, 2009.

E.J. Candes and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse problems*, 23: 969, 2007.

K.L. Chung. *Markov Chains with Stationary Transition Probabilities*. Springer, 1967.

T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

J.L. Doob. *Stochastic Processes*. Wiley, 1953.

D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. URL http://www.cs.cornell.edu/home/kleinber/networks-book/.

R.G. Gallager. *Low Density Parity Check Codes*. M.I.T. Press, 1963.

R.G. Gallager. *Stochastic Processes: Theory for Applications*. Cambridge University Press, 2014.

G.C. Goodwin and K.S. Sin. *Adaptive Filtering Prediction and Control*. Dover, 2009.

G.R. Grimmett and D.R. Stirzaker. *Probability and Random Processes*. Oxford University Press, 2001.

B. Hajek and T. Van Loon. Decentralized dynamic control of a multiple access broadcast channel. *IEEE Transactions on Automatic Control*, AC-27(3):559–569, 1982.

M. Harchol-Balter. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press, 2013.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer-Verlag, 2009.

D.A. Huffman. A Methd for the construction of minimum-redundancy codes. *Proceeding of the IRE*, pages 1098–1101, 1952.

R.E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

F. Kelly. *Reversibility and Stochastic Networks*. Wiley, 1979.

F. Kelly and E. Yudovina. *Lecture Notes on Stochastic Networks*. 2013. URL http://www.statslab.cam.ac.uk/~frank/STOCHNET/LNSN/book.pdf.

L. Kleinrock. *Queueing Systems v.1 and v.2*. J. Wiley, 1975-6.

P.R. Kumar and P.P. Varaiya. *Stochastic Systems: Estimation, Identification and Adaptive Control*. Prentice- Hall, 1986.

E.L. Lehmann. *Testing Statistical Hypotheses, 3rd* Edition. Springer, 2010.

J. D. C. Little. A Proof for the Queuing Formula: $L = \lambda W$. *Operations Research*, 9(3):383–401, 1961.

S.G. Mallat and Z. Zhang. Matching Pursuits with Time-Frequency Dictionaries. *IEEE Transactions on Signal Processing*, pages 3397–3415, 1993.

J. Neveu. *Discrete Parmeter Martingales*. North-Holland / American Elsevier, 1975.

J; Neyman and E. S. Pearson. On the Problem of the most Efficient Tests of Statistical Hypotheses. *Phil. Trans. R. Soc. Lond.*, pages 289–337, 1933.

L. Page. Method for node ranking in a linked database. *United States Patent and Trademark Office*, 6,285,999, 2001.

J. Pitman. *Probability*. Springer Texts in Statistics, 1993.

J. Proakis. *Digital Communications, 4th Edition*. McGraw-Hill Science/Engineering/Math, 2000.

T. Richardson and R. Urbanke. *Modern Coding Theory*. Cambridge University Press, 2008.

E. Roche. EM algorithm and variants: an informal tutorial. 2012. URL arXiv:1105.1476v2[stat.CO].

S.M. Ross. *Introduction to Stochastic Dynamic Programming*. Academic Press, 1995.

R. Srikant and L. Ying. *Communication Networks: An Optimization, Control, and Stochastic Networks Perspective*. Cambridge University Press, 2014.

G. Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 2009.

E.L. Strehl and M.L. Littman. Online linear regression and its application to model-based reinforcement learning. In *In Advances in Neural Information Processing Systems 20 (NIPS-07*, pages 737–744, 2007.

D. Tse and P. Viswanath. *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

M.J. Wainwright and M. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers, 2008.

J. Walrand. *An Introduction to Queueing Networks*. Prentice-Hall, 1988.

E. Wong and B. Hajek. *Stochastic Processes in Engineering Systems*. Springer-Verlag, 1985.

“这本书适应了信息时代的需要，是EECS（电子工程和计算机科学）专业学生学习概率论的终极选择。作者在EECS和概率论这两个领域都有极高的造诣，他没有按照常规的教材顺序，而是以具体的工程问题作为切入点来介绍基本的概率论知识。不论是作为课程的教材还是作为补充阅读材料，这都是一本无与伦比的好书。”

—— Vivek S. Borkar

印度孟买塔塔基础研究院技术与计算机科学学院资深教授

☆ 囊括当前热门应用 ☆

PageRank、多路复用、数字链路、追踪定位、语音识别、路线规划……

☆ 剖析概率论基本概念 ☆

大数定律、中心极限定理、高斯分布、特征方程、似然比、线性最小平方估计、条件期望、期望最大化……

☆ 使用Matlab进行仿真和计算 ☆

矩阵、随机变量、马尔可夫链、置信区间、数字链路、LDPC码、卡尔曼滤波器……

图灵社区：iTuring.cn
新浪微博：@图灵教育 @图灵社区 @图灵新知
微信公众号：turingbooks ituring_interview
反馈/投稿/推荐邮箱：contact@turingbook.com
读者热线：(010) 51095186-600

分类建议 数学/概率论

人民邮电出版社网址：www.ptpress.com.cn

ISBN 978-7-115-39896-3

ISBN 978-7-115-39896-3
定价：69.00元

看完了

如果您对本书内容有疑问，可发邮件至contact@turingbook.com，会有编辑或作译者协助答疑。也可访问图灵社区，参与本书讨论。

如果是有关电子书的建议或问题，请联系专用客服邮箱：ebook@turingbook.com。

在这里可以找到我们：

微博 @图灵教育 : 好书、活动每日播报
微博 @图灵社区 : 电子书和好文章的消息
微博 @图灵新知 : 图灵教育的科普小组
微信 图灵访谈 : ituring_interview，讲述码农精彩人生
微信 图灵教育 : turingbooks